# ON THE EVOLUTION OF CONTINUED FRACTIONS IN A FIXED QUADRATIC FIELD

MENNY AKA AND URI SHAPIRA

Abstract. We prove that the statistics of the period of the continued fraction expansion of certain sequences of quadratic irrationals from a fixed quadratic field approach the 'normal' statistics given by the Gauss-Kuzmin measure. As far as we know, these are the first non-average results about the statistics of the periods of quadratic irrationals. As a by-product, the growth rate of the period is analyzed and, for example, it is shown that for a fixed integer $k$ and a quadratic irrational $\alpha$, the length of the period of the continued fraction expansion of $k^n\alpha$ equals $c'k^n + o(k^{(1-\frac{1}{16})n})$ for some positive constant $c'$. This improves results of Lagarias and Grisel. The results are derived from the main theorem of the paper, which establishes an equidistribution result regarding single periodic geodesics along certain paths in the Hecke graph. The results are effective and give rates of convergence and the main tools are spectral gap (effective decay of matrix coefficients) and dynamical analysis on $S$-arithmetic homogeneous spaces.

## 1. Introduction

### 1.1. **Continued fractions.**

The elementary theory of continued fractions starts by assigning to each real number $x \in [0,1]\setminus\mathbb{Q}$ an infinite sequence of positive integers[1][2] referred to as the *continued fraction expansion* of $x$ (abbreviated c.f.e); namely to each number $x$ corresponds a sequence $\{a_n(x)\}_{n\in\mathbb{N}}$ which is characterized by the requirement

$$x = \lim_{n\to\infty} \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots \frac{1}{a_n}}}} \tag{1.1}$$

We refer to the numbers $a_n(x)$ as the digits of the expansion as in (1.1). When $x$ is understood we usually write $a_i$ for the $i$'th digit of the c.f.e of $x$.

Given a number $x$, it is natural to ask for information regarding the statistical properties of its c.f.e; that is, for any finite sequence of natural numbers $w = (w_1, \ldots, w_k)$ (referred to hereafter as a *pattern*) one is interested in the *frequency of appearance* of the pattern $w$ in the c.f.e of $x$, or in other words in the value of the limit[3]

$$D(x,w) = \lim_N \frac{1}{N}\#\left\{1 \le n \le N : w = (a_{n+1}, \ldots, a_{n+k})\right\}. \tag{1.2}$$

[1] We shall completely ignore the rational numbers, which correspond to finite sequences as well as real numbers outside the unit interval, for which an additional integer digit $a_0$ is needed.

[2] This correspondence is in fact a homeomorphism when $\mathbb{N}^{\mathbb{N}}$ is considered with the product topology.

[3] The limit does not always exist.

It turns out (as will be explained shortly) that this frequency exists and equals some explicit integral (depending only on $w$) for Lebesgue almost any $x$.

To see this, note that the c.f.e correspondence $x \leftrightarrow \{a_n(x)\}$ fits in the commutative diagram

$$\begin{tikzcd}
\mathbb{N}^{\mathbb{N}} \arrow[r, "\sigma"] \arrow[d] & \mathbb{N}^{\mathbb{N}} \arrow[d] \\
{[0,1] \setminus \mathbb{Q}} \arrow[r, "S"] & {[0,1] \setminus \mathbb{Q}} ,
\end{tikzcd} \tag{1.3}$$

where $S(x) = \{\frac{1}{x}\} = \frac{1}{x} - \lfloor \frac{1}{x} \rfloor$ is the so-called Gauss map and $\sigma$ is the shift map $\sigma(a_1, a_2, \dots) = (a_2, a_3, \dots)$. Gauss observed in 1845 that $S$ preserves the measure given by

$$\nu(A) = \frac{1}{\log(2)} \int_A \frac{1}{1+x} dx, \tag{1.4}$$

on the unit interval (this is the so called Gauss-Kuzmin measure). The map $S$ is ergodic with respect to $\nu$ which implies by the pointwise ergodic theorem (see for example [EW11, §2.6,§9.6]) that for $\nu$ (or equivalently Lebesgue) almost any $x$ and any pattern $w = (w_1, \dots, w_k)$, the frequency $D(x, w)$ defined in (1.2) exists. More precisely, if we let

$$I_w = \{x \in [0,1] \setminus \mathbb{Q} : w = (a_1(x), \dots, a_k(x))\} \tag{1.5}$$

denote the interval consisting of those points for which the c.f.e starts with the pattern $w$, then the pointwise ergodic theorem tells us that the ergodic averages of the characteristic function of $I_w$ converge almost surely to $\nu(I_w)$; that is

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{N-1} \chi_{I_w}(S^i(x)) = \nu(I_w), \tag{1.6}$$

for $\nu$ almost any $x$. As the set of possible patterns is countable we conclude that for Lebesgue almost any $x$ (1.6) holds for any pattern $w$. It is straightforward to check using the commutation in (1.3) that the limit in (1.6) is equal to the limit in (1.2).

1.2. **Quadratic irrationals.** By Lagrange's Theorem (see for example [EW11, §3.3]) the numbers $x$ for which the c.f.e is eventually periodic are exactly the quadratic irrationals; that is real numbers which are roots of irreducible quadratic polynomials over the rationals. For quadratic irrationals (which clearly form a Lebesgue-null set) it is clear that the limit in (1.2) always exists and is different from the almost sure value of the frequency.

1.3. **General goal.** In this paper we investigate the behavior of $D(x, w)$ where $x$ varies in some fixed quadratic field. We make the convention to consider $x \bmod 1$ instead of $x$. This influences only the 0'th digit in the classical discussion on continued fractions and does not effect of course any statistical property of the c.f.e. Our approach manages to deal with sequences $x_n$ which are related arithmetically in a way that involves only finitely many primes (see for example Theorems 2.1,2.7 and Remark 2.6). Nevertheless, the discussion leaves many natural open questions, a few of which we state below (see §2.7).

The following theorem demonstrates the flavor of our results, which will be stated below in increasing levels of generality, culminating in Theorem 4.9.

**Notation 1.1.** Throughout this paper we use the notation $\ll$ in the following manner: Given two quantities $A, B$ depending on some set of parameters $P$, we denote $A \ll B$ if there exists some absolute constant $c > 0$ (independent of any varying parameter) such that $A \leq cB$. Given a subset $P'$ of the parameter set $P$, we denote $A \ll_{P'} B$ if there exists a constant $c_{P'} > 0$, depending possibly on the parameters in $P'$, such that $A \leq c_{P'}B$.

We denote by $|I_w|$ the length of the interval $I_w$ defined in (1.5). The following theorem follows from Corollary 2.2 as explained in Remark 2.3.

**Theorem 1.2.** *Let $\alpha$ be a quadratic irrational, $w = (w_1, \ldots, w_k)$ a finite pattern of natural numbers, and $p$ a prime number. We have that $D(p^n\alpha, w) \to \nu(I_w)$ as $|n| \to \infty$. Moreover, the following effective estimate holds*

$$|D(p^n\alpha, w) - \nu(I_w)| \ll_{\alpha,p} |I_w|^{-1} p^{-\frac{|n|}{32}}.$$

1.4. **The measures $\nu_\alpha$.** Before ending this short introduction we adopt a slightly different viewpoint which will be more convenient to our discussion. As the c.f.e of a quadratic irrational is eventually periodic, it follows from (1.3) that the orbit $\{S^n\alpha\}_{n\in\mathbb{N}}$ of $\alpha$ under the Gauss map, is eventually periodic with some period $P_\alpha = \{x_1, \ldots, x_\ell\} \subset [0,1] \setminus \mathbb{Q}$ with $S(x_i) = x_{i+1}$ for all $i < \ell$ and $S(x_\ell) = x_1$. Let us denote by $\nu_\alpha$ the normalized counting measure on the set $P_\alpha$. Note that with this notation $D(\alpha, w) = \nu_\alpha(I_w)$, and so Theorem 1.2 above could be restated as saying that $\nu_{p^n\alpha}$ converge in the weak$^*$ topology to the Gauss-Kuzmin measure $\nu$ (with an explicit estimate on the rate of convergence). All convergence statements regarding measures will refer to the weak$^*$ topology. We sometime say that a sequence of measures $\mu_n$ equidistributes to a measure $\mu$ to mean that it converges to it.

## 2. Results

The results appearing in this paper divide into two. One portion deals with effective equidistribution of single periodic geodesics arising from a fixed quadratic extension of $\mathbb{Q}$ and the other portion deals with harvesting the conclusions of the first part in the theory of continued fractions. Though in order to describe the results regarding continued fractions we do not need too much preparation, the statement of Theorem 4.9, which is the main result of this paper and belongs to the first portion, needs some preparations. We therefore describe in this section the precise results regarding continued fractions, but make the compromise to describe the results regarding periodic geodesics in a more ‘hands on’ way and refer the reader to §4 for the more accurate state of affairs.

2.1. **Height and support.** Given a rational number $q = \frac{k_1}{k_2}$, where $k_1, k_2$ are co-prime integers, we define the *height* of $q$ to be

$$\mathrm{ht}(q) = k_1k_2. \tag{2.1}$$

Given a set of primes $S$, we say that $q$ is *supported on* $S$ if all the primes dividing $k_1, k_2$ are in $S$.

2.2. **The exponent $\delta_0$.** Appearing in the results below is a parameter $\frac{25}{64} \leq \delta_0 \leq \frac{1}{2}$ whose exact value is not known (although according to the Ramanujan conjecture $\delta_0 = \frac{1}{2}$). The bigger it is the stronger the statements are and the best known lower bound for it to this date is $\delta_0 \geq \frac{25}{64}$; a bound given by Kim and Sarnak in the appendix of [Kim03]. The meaning of this parameter will be explained in §6.3 but for the moment we will simply mention that it has to do with certain complementary series representations of $\mathrm{GL}_2$ not appearing in the spectral decomposition of some unitary representation.

2.3. **Results regarding continued fractions.** To the best of our knowledge, there are basically no results in the literature regarding the statistical evolution of the period of the c.f.e of quadratics (not on average) which is the subject of the results in this section. We note though that there is plenty of literature regarding the evolution of the length of the period (see §2.6 and Corollary 2.5). The following results are proved in §8 and all of them are consequences of Theorem 4.9. Recall that a function $f : Y \to Z$ between metric spaces $(Y, \mathrm{d}_Y), (Z, \mathrm{d}_Z)$ is said to be *$\kappa$-Lipschitz*, for $\kappa \geq 0$, if for any $y_1, y_2 \in Y$ we have $\mathrm{d}_Z(f(y_1), f(y_2)) \leq \kappa\, \mathrm{d}_Y(y_1, y_2)$. We then say that $\kappa$ is a Lipschitz constant for $f$.

**Theorem 2.1.** *Let $\alpha$ be a quadratic irrational, $S$ a finite set of primes, $q$ a rational number supported on $S$, and $\epsilon > 0$. For any $\kappa$-Lipschitz function $f : [0,1] \to \mathbb{C}$ the following holds*

$$\left| \int_0^1 f d\nu - \int_0^1 f d\nu_{q\alpha} \right| \ll_{\alpha,S,\epsilon} \max\left\{ \|f\|_\infty, \kappa \right\} \mathrm{ht}(q)^{-\frac{\delta_0}{6}+\epsilon}, \tag{2.2}$$

*where* $\|f\|_\infty = \sup\{|f(t)| : t \in [0,1]\}$.

When we use Theorem 2.1 to try and estimate the frequency of a pattern in the period of the c.f.e of $q \cdot \alpha$ we obtain the following corollary which we leave without proof.

**Corollary 2.2.** *Let $\alpha, S, q, \epsilon$ be as in Theorem 2.1. For any finite pattern $w = w_1 \dots w_k$ of digits,*

$$|D(q\alpha, w) - \nu(I_w)| \ll_{\alpha,S,\epsilon} |I_w|^{-1} \mathrm{ht}(q)^{-\frac{\delta_0}{12}+\epsilon}. \tag{2.3}$$

The exponent in (2.2) is cut in half in (2.3) as a result of the fact that $\chi_{I_w}$ is not Lipschitz and one needs to use an approximation of it in order to apply Theorem 2.1.

**Remark 2.3.** Theorem 1.2 is obtained from the above corollary by taking $q = p^n$, the Kim-Sarnak exponent $\delta_0 = \frac{25}{64}$, and choosing $\epsilon = \frac{1}{768}$ so that $-\frac{\delta_0}{12} + \epsilon = -\frac{1}{32}$.

In particular, it follows from Theorem 2.1 that if $q_n$ is a sequence of rationals supported on $S$ with $\mathrm{ht}(q_n) \to \infty$, then $\nu_{q_n\alpha}$ equidistributes to the Gauss-Kuzmin measure $\nu$. The following example which was essentially communicated to us by A. Ubis shows that one cannot expect such convergence to hold for a general sequence of rationals $q_n$ with $\mathrm{ht}(q_n) \to \infty$.

**Example 2.4.** Let $D$ be a fundamental discriminant such that the negative Pell equation $x^2 - Dy^2 = -1$ has an integer solution (see [Lag80],[FK10] for example). Whenever the equation is soluble, it corresponds to a fundamental unit $\epsilon_D = k_1 + n_1\sqrt{D}$ in $\mathbb{Q}(\sqrt{D})$ of norm $-1$ and in turn, the odd powers $\epsilon_D^j$ correspond to infinitely many further solutions of the negative Pell equation $(k_j, n_j)$. Fix such $D$ and for odd $j$ let $\alpha_j$ solve the equation $x = 2k_j + \frac{1}{x}$. That is the c.f.e of $\alpha_j$ is purely periodic with period of length 1 of digit $2k_j$ (note that here we abuse the notation introduced above and we do record the 0 digit). Solving for $x$ in the above equation we see that $\alpha_j$ could be chosen to be $k_j + \sqrt{k_j^2 + 1}$. As $(k_j, n_j)$ solve the negative Pell equation for $D$ we get $\alpha_j = k_j + n_j\sqrt{D}$ which shows that the measures $\nu_{n_j\sqrt{D}}$ are not converging to the Gauss-Kuzmin measure and in fact are atomic measures supported on single points.

The following corollary discusses the growth rate of the period. It follows from the argument yielding Theorem 2.1 and is proved at the end of §8. For a quadratic irrational $\alpha$ we denote by $|P_\alpha|$ the cardinality of the support of $\nu_\alpha$. Note that $|P_\alpha|$ is the *length of the period of the c.f.e of* $\alpha$. Apparently, it was known already to Dirichlet that the length of the period of the c.f.e of $5^n\sqrt{5}$ grows very quickly. In the Appendix of [Lag80], using the methods of Dirichlet, Lagarias shows that under some restrictive assumptions on $\alpha$ one has that for any integer $k$ there exists a constant $C$ for which, $C\frac{k^n}{n} < |P_{k^n\alpha}|$. Under some restrictive assumptions on $\alpha$, Grisel [Gri98] proved a stronger estimate of the form $C_1 k^n \leq |P_{k^n\alpha}| \leq C_2 k^n$. The following corollary strengthens these results in several respects. Without any restrictive assumption on $\alpha$ we establish that for a fixed $k$ $|P_{k^n\alpha}| = c'k^n + o(k^{(1-\delta)n})$, where the constant $c' > 0$ depend on $\alpha$ and $k$, and $\delta$ could be taken to be $\frac{1}{16}$ similarly to Remark 2.3.

**Corollary 2.5.** *Let $\alpha, S, q,$ and $\epsilon$ be as in Theorem 2.1. There exists an absolute positive constant $c$ and and a positive function $c_\alpha(q)$ which attains only finitely many values (as $q$ ranges over the rationals supported on $S$) such that*

$$\left| c_\alpha(q)\frac{|P_{q\alpha}|}{\mathrm{ht}(q)} - c \right| \ll_{\alpha,S,\epsilon} \mathrm{ht}(q)^{-\frac{\delta_0}{6}+\epsilon}. \tag{2.4}$$

*Moreover, if $q_n$ satisfies $\mathrm{ht}(q_n) | \mathrm{ht}(q_{n+1})$ (for example $q_n = k^n$), then $c_\alpha(q_n)$ is constant for large values of $n$.*

**Remark 2.6.** Consider for simplicity the case where $S = \{p\}$ consists of a single prime. In the above theorems we considered the two sequences $p^{\pm n}\alpha$, where $\alpha$ is a quadratic irrational. In fact, as explained in Remark 4.14, corresponding results hold for other sequences. As an example, given a sequence $\{j_i\}_{i=0}^{\infty}$, where $j_i \in \{0, \ldots, p-1\}$, if we define recursively $\alpha_{-1} = \alpha$ and $\alpha_{n+1} = \frac{\alpha_n + j_n}{p}$ (i.e. $\alpha_n = p^{-n}(\alpha + \sum_{i=0}^{n-1} j_i p^i)$), then the following holds: If $p$ does not split in $\mathbb{Q}(\alpha)$, then the estimates (2.2), (2.3), and (2.4) with $q\alpha$ replaced with $\alpha_n$ and $\mathrm{ht}(q)$ replaced by $p^n$ still hold. If $p$ splits in $\mathbb{Q}(\alpha)$ a more restrictive statement holds: If the sequence $\{j_i\}$ used to define $\alpha_n$ is eventually periodic then the estimates (2.2), (2.3), and (2.4) with $q\alpha$ replaced with $\alpha_n$ and $\mathrm{ht}(q)$ replaced by $p^n$ still hold, but the implicit constants depend on the sequence $\{j_i\}$.

The following amusing phenomenon comes out of our analysis. Recall that the group $\mathrm{SL}_2(\mathbb{Z})$ acts on the real line by Möbius transformations. It is well known that for $\alpha \in \mathbb{R}$, the orbit $\{\gamma \cdot \alpha : \gamma \in \mathrm{SL}_2(\mathbb{Z})\}$ is characterized as the set of all real numbers whose c.f.e has the same tail as the c.f.e of $\alpha$.

**Theorem 2.7.** *Let $\alpha, S$ be as in Theorem 2.1. The implicit constants appearing in all the results of §2.3 (namely in* (2.2), (2.3), *and* (2.4)*) may be taken to be uniform for all the quadratic irrationals in $\{\gamma \cdot \alpha : \gamma \in \mathrm{GL}_2(\mathbb{Z})\}$ if and only if all the primes in $S$ do not split in the quadratic extension $\mathbb{Q}(\alpha)$.*

2.4. **Results regarding periodic geodesics.** The tight connection between the theory of continued fractions and the geodesic flow on the unit tangent bundle to the modular surface is by now considered classical and dates back to E. Artin's 1924 paper [Art82]. The proofs of the above theorems utilizes this connection and follows from a corresponding equidistribution theorem of certain geodesic loops in this space. Nonetheless, quite a bit of technical work is needed (see §8,9) to derive the above effective results regarding continued fractions from Theorems 2.8 below.

For any quadratic irrational $\alpha$, let $\alpha'$ denote its Galois conjugate and let us define

$$g_\alpha = \begin{pmatrix} \alpha & \alpha' \\ 1 & 1 \end{pmatrix} \text{ if } \alpha - \alpha' > 0, \text{ and } g_\alpha = \begin{pmatrix} \alpha & -\alpha' \\ 1 & -1 \end{pmatrix} \text{ otherwise.} \tag{2.5}$$

The homogeneous space[4]

$$X_\infty = \mathrm{PGL}_2(\mathbb{Z}) \backslash \mathrm{PGL}_2(\mathbb{R})$$

is naturally identified with the unit tangent bundle to the modular surface[5] (the subscript $\infty$ will become clear in §3). Via this identification the geodesic flow corresponds to the action from the right of the diagonal group $\{a_\infty(t) : t \in \mathbb{R}\}$, where $a_\infty(t) = \mathrm{diag}\,(e^t, 1) \in \mathrm{PGL}_2(\mathbb{R})$. By Lemma 4.4 below, the point $x_\alpha = \mathrm{PGL}_2(\mathbb{Z})g_\alpha \in X_\infty$ is a periodic point for the geodesic flow; let $\mu_\alpha$ denote the normalized length measure supported on the geodesic loop through $x_\alpha$. Finally, let $m_\infty$ denote the $\mathrm{PGL}_2(\mathbb{R})$-invariant probability measure on $X_\infty$. The following theorem establishes, for example, the equidistribution $\mu_{3^n\sqrt{2}} \to m_\infty$.

**Theorem 2.8.** *Let $\alpha$ be a quadratic irrational and $S$ a finite set of primes, $q$ a rational number supported on $S$, and $\epsilon > 0$. For any $\kappa$-Lipschitz function $\varphi \in L^2(X_\infty, m_S)$ one has*

$$\left| \int \varphi d\mu_{q\alpha} - \int \varphi dm_\infty \right| \ll_{\alpha,S,\epsilon} \max\left\{ \|\varphi\|_2, \kappa \right\} \mathrm{ht}(q)^{-\frac{\delta_0}{2}+\epsilon}. \tag{2.6}$$

In fact, as explained in Remark 4.10, Theorem 2.8 follows from the more general Theorem 4.9. In a nutshell, Theorem 4.9 asserts that if one considers the $p$-Hecke tree through $x_\alpha$, then unless an obvious obstacle is present, when taking a sequence $x_n$ on the tree which drifts away from the root $x_\alpha$, the periodic geodesics through $x_n$ must equidistribute to $m_\infty$. Moreover, the amusing phenomena referred to in Theorem 2.7 asserts that this

[4] This does not reflect the authors preference saying that the right action is the left one.

[5] See §8.

equidistribution is uniform in the distance from $x_n$ to the root if and only if the prime $p$ do not split in the corresponding quadratic extension of $\mathbb{Q}$. For a richer family of measures than $\mu_{q\alpha}$ for which an estimate such as (2.6) holds (in the spirit of Remark 2.6) we refer the reader to Remark 4.14.

2.5. **About the paper.** The first version of this paper was much shorter and elementary but it only established the equidistribution $\mu_{p^n\alpha} \to m_\infty$ (and its implication $\nu_{p^n\alpha} \to \nu$) and lacked the effective nature that appears in the results presented here. The decision to write the current version has a few serious disadvantages. The first is that the underlying simple argument is somewhat hidden behind many technical preparations which are needed for the effective results (and thus making the paper longer), and the second is that it narrows down the readership, as sophisticated tools such as effective decay of matrix coefficients are used. In order to remedy this, we plan on writing a survey paper [AS], in which only the 'soft' convergence $\mu_{p^n\alpha} \to m_\infty$ is proved. We expect this survey to be essentially self contained and approachable for any advanced graduate student. It should also serve the purpose of preparing readers with less background for reading the current paper.

2.6. **References to existing results.** Although the question of the evolution of the c.f.e along arithmetically defined sequences in a fixed quadratic field is extremely natural, we did not find too many relevant papers to cite. Some earlier works studying the statistics of the period 'in average' (and also not in a fixed field), were initiated by Arnold (see [Arn08],[Arn07],[Ler10] and the references therein). See also [Pol86]. Other works, mostly related to the length of the period, which the reader might find related, may be found for example in many of the papers of Golubeva (such as [Gol02]) and in [Gri98],[BL05][MF93],[CZ04],[Coh77],[Hic73],[Kei]. Standing out in this context is the recent paper of McMullen which provides examples of sequences of quadratic irrationals in a fixed quadratic field with uniformly bounded c.f.e digits [McM09]. We suspect that it should be very interesting to compare in detail how McMullen's results fit together with the results of the present paper.

As for results regarding periodic geodesics the situation is completely different and we will not try to list below all the relevant earlier work that has been done in the subject. We do wish to comment though, that as will be explained in §5, Theorems 2.8,4.9 are closely related to the works of Benoist and Oh [BO07] and to Duke's Theorem [Duk88],[ELMV]. In fact, the non-effective equidistribution $\mu_{q_n\alpha} \to m_\infty$ (where the $q_n$ are supported on $S$ and $\mathrm{ht}(q_n) \to \infty$) follows from [BO07, Theorem 1.1] (and from Duke's Theorem) by a short elementary argument as will be explained in §5.1. Both the works of Benoist and Oh, and Duke's Theorem deal with the equidistribution of certain collections of closed geodesics and the phenomenon that happens in our case is that the single geodesic corresponding to $q\alpha$ occupies a positive proportion of the collection (see Lemma 5.1).

Although the non-effective equidistribution $\mu_{q_n\alpha} \to m_\infty$ can be deduced from known results as stated above, our argument is independent of them and moreover, as noted above, it may be adopted to give an essentially self contained proof of the non-effective result [AS].

2.7. **Some open problems.** We list below a few questions which emerge from our discussion and remain unsolved. Each of the problems below have a corresponding problem stated in terms of periodic geodesics on the modular surface.

(1) Give satisfactory sufficient conditions on a sequence of rationals $q_n$ to ensure that for a quadratic irrational $\alpha$, the sequence of measures $\nu_{q_n\alpha}$ equidistribute to the Gauss-Kuzmin measure $\nu$. It might be interesting to replace the quantifiers and allow the conditions to depend on $\alpha$.
(2) Is it true that for a quadratic irrational $\alpha$ which is not a unit in the ring of integers of $\mathbb{Q}(\alpha)$, the sequence of measures $\nu_{\alpha^n}$ always equidistribute to the Gauss-Kuzmin measure along the subsequence of $n$'s for which $\alpha^n$ is irrational (see [CZ04]). Note that our results deal with the case $\alpha = \sqrt{d}$.
(3) Let $p_n$ be an enumeration of the primes. Are there any quadratic irrationals $\alpha$ for which $\nu_{p_n\alpha}$ equidistribute to the Gauss-Kuzmin measure.
(4) Is it true that for any quadratic irrational $\alpha$ there exist a sequence of distinct primes $p_n$ so that $\nu_{p_n\alpha}$ equidistribute to the Gauss-Kuzmin measure.

2.8. **Acknowledgments.** We benefited from numerous stimulating conversations with Manfred Einsiedler, Elon Lindenstraus and various other people which we try to name below. We would like to express our gratitude and appreciation to them for the generosity in which they exchange valuable ideas and create an ‘open source’ healthy environment in our community doing Homogeneous Dynamics. Thanks are also due to Yann Bugeaud, Étienne Fouvry, Tsachik Gelander, Alex Gorodnik, Chen Meiri, Philippe Michel, Shahar Mozes, Hee Oh, Peter Sarnak, Adrian Ubis, Akshay Venkatesh, and Barak Weiss.

Both Authors enjoyed the warm hospitality of the Centre Interfacultaire Bernoulli and the GANT semester held there.

M.A acknowledges the support of the Advanced research Grant 226135 from the European Research Council, the ISEF foundation, the Ilan and Asaf Ramon memorial foundation, and the Hoffman Leadership and Responsibility fellowship program at the Hebrew University of Jerusalem.

U.S acknowledges the support of the Advanced research Grant 228304 from the European Research Council.

## 3. Preliminaries

3.1. **Notation.** For a prime $p$ we let $\mathbb{Q}_p$ denote the field of p-adic numbers and by $\mathbb{Z}_p$ the ring of $p$-adic integers. We sometimes denote $\mathbb{Q}_\infty = \mathbb{R}$. The set $\mathbf{P} = \{\infty\} \cup \{p \in \mathbb{N} : p \text{ is a prime}\}$ will be referred to as the set of *places* of $\mathbb{Q}$ – the primes being the *finite places*.

Let $S \subset \mathbf{P}$ be given. Throughout we let $S_f = S \setminus \{\infty\}$. We denote by $\mathbb{Q}_S, \mathbb{Z}_S$ the product rings $\prod_{v\in S} \mathbb{Q}_v, \prod_{v\in S} \mathbb{Z}_v$ respectively (the latter makes sense only when $\infty \notin S$). Let $\mathbb{G}$ denote either one of the algebraic groups $\mathrm{PGL}_2, \mathrm{PSL}_2$. We denote $G_S = \mathbb{G}(\mathbb{Q}_S)$. We denote an element $g \in G_S$ by a sequence $g = (g_v)_{v\in S}$ where $g_v$ is a $2\times 2$ matrix over $\mathbb{Q}_v$ (note the slight abuse of notation). If $\infty \in S$ we usually abbreviate and write $g = (g_\infty, g_f)$

where $g_f$ denotes the tuple of the components corresponding to the finite places in $S$. The identity elements in the various groups are denoted by $e$ with the corresponding subscript. Thus for example $e_S = (e_\infty, e_f)$.

We may view the group $\Gamma_S = \mathbb{G}(\mathbb{Z}[\frac{1}{p} : p \in S_f])$ as a subgroup of $G_S$ (embedded diagonally). If $\infty \in S$, it is well known that $\Gamma_S$ is a lattice in $G_S$. We denote by $X_S$ the homogeneous space $\Gamma_S \backslash G_S$ and by $m_S$ the $G_S$-invariant probability measure on it. The real quotient $X_\infty = \Gamma_\infty \backslash G_\infty$ is a *factor* of $X_S$ in a natural way: Let $K_\infty$ denote the maximal compact subgroup of $G_\infty$ which is the (projective) orthogonal group $\mathrm{PO}_2(\mathbb{R})$ in the case of $\mathrm{PGL}_2$ or $\mathrm{PSO}_2$ in the case of $\mathrm{PSL}_2$. For a finite place $p \in \mathbf{P}_f$ we let $K_p = \mathbb{G}(\mathbb{Z}_p)$. We then let $K_S$ denote the product $\prod_{v \in S} K_v$. If $\infty \in S$, the double coset space $X_S / K_{S_f} = \Gamma_S \backslash G_S / K_{S_f}$ is naturally identified with $X_\infty$. We denote by $\pi : X_S \to X_\infty$ the natural projection.

**Remark 3.1.** In practice, given $x = \Gamma_S(g_\infty, g_f) \in X_S$ with representative $(g_\infty, g_f)$ such that $g_f \in K_{S_f}$, the projection $\pi(x)$ is defined to be $\Gamma_\infty g_\infty$. In other words, $\pi^{-1}(\Gamma_\infty g_\infty) = \left\{\Gamma_S(g_\infty, g_f) : g_f \in K_{S_f}\right\}$. Another useful observation to keep in mind here is that two points $x_1 = \Gamma_S(g_\infty, g_f), x_2 = \Gamma_S(g_\infty, h_f)$ are in the same fiber (that is $\pi(x_1) = \pi(x_2)$) if and only if the quotient $g_f^{-1} h_f$ belongs to $K_{S_f}$.

The group $G_S$ (and all its subgroups) act on $X_S$ by right translation. In particular, if $T \subset S$, we may view $G_T$ (and its subgroups) as a subgroup of $G_S$ and thus it acts on $X_S$. Note that $\pi : X_S \to X_\infty$ intertwines the $G_\infty$ actions. Of particular interest to us will be the action of the real diagonal group $A_\infty = \{\mathrm{diag}\,(e^t, 1) : t \in \mathbb{R}\}$, the elements of which we often write as $a_\infty(t) = \mathrm{diag}\,(e^t, 1)$.

We say that an orbit $xL$ of a closed subgroup $L < G_S$ through a point $x \in X_S$ is *periodic* if it supports an $L$-invariant probability measure. Such a measure is unique and we refer to it as the Haar measure on the periodic orbit. Compact orbits are always periodic. Given a measure $\mu$ on $X_S$ and $g \in G_S$ we let $g_*\mu$ denote the pushed forward measure by right translation by $g$. This notation is a bit awkward as $(gh)_*\mu = h_*(g_*\mu)$. This will not bother us as we will only use commutative subgroups to push measures.

The Lie algebra of $G_v$ will be denoted by $\mathfrak{g}_v$ and is naturally identified with the space of traceless $2 \times 2$ matrices over $\mathbb{Q}_v$. Similarly to the notation introduced above we will denote by $\mathfrak{g}_S = \oplus_{v \in S} \mathfrak{g}_v$ the Lie algebra of $G_S$. A basic fact that we will use is that if $S$ is finite and $L < G_S$ is a closed subgroup then $L$ contains an open product subgroup $\prod_{v \in S} L_v$ which allows us to speak of the Lie algebra of $L$ which will be denoted $\mathrm{Lie}\,(L)$. The exponential map $\exp_v : \mathfrak{g}_v \to G_v$ is defined for any place $v$ by the usual power series and in fact, is only well defined for finite places on a certain neighborhood of 0. We denote its inverse by $\log_v$ (it is defined on a small enough neighborhood of $e_v$) and use the obvious notation $\exp_S, \log_S$ to denote the corresponding maps from the corresponding domains in $\mathfrak{g}_S, G_S$ respectively.

Given an element $g \in G_S$ and an element $u$ (either of $G_S$ or of $\mathfrak{g}_S$), we denote by $u^g$ the conjugation $g^{-1}ug$. If $g$ is semisimple we denote by $(\mathfrak{g}_S)_g^{\mathrm{ws}}$ the *weak stable* subalgebra of $\mathfrak{g}_S$. It is defined as the direct sum of the eigenspaces (of the operator $u \mapsto u^g$) of modulus

$\leq 1$ or equivalently

$$(\mathfrak{g}_S)^{\mathrm{ws}}_g = \left\{ u \in \mathfrak{g}_S : \left\{ u^{(g^n)} \right\}_{n>0} \text{ is bounded in } \mathfrak{g}_S \right\}.$$

For each place $v$ we equip $G_v, \mathfrak{g}_v$ with metrics in the following way: For $v = \infty$ we start with an inner product on $\mathfrak{g}_\infty$ which is right $K_\infty$-invariant and use left translation to make it into a left invariant Riemanian metric on $G_\infty$ which is also right $K_\infty$-invariant. This Riemannian metric induces left $G_\infty$-invariant, bi-$K_\infty$-invariant metric on $G_\infty$. For a finite place $v$, we start with a bi-$K_v$-invariant metric $\mathrm{d}_{K_v}$ on $K_v$ (such that $K_v$ equals the closed unit ball around $e_v$) and make it into a left invariant metric on $G_v$ (which is also right $K_v$-invariant) by setting $\mathrm{d}_{G_v}(g_1, g_2) = 2$ if $g_1^{-1} g_2 \notin K_v$ and $\mathrm{d}_{G_v}(g_1, g_2) = \mathrm{d}_{K_v}(g_1^{-1} g_2, e_v)$. On the Lie algebra $\mathfrak{g}_v$ we take the metric given by $\mathrm{d}_{\mathfrak{g}_v}(u, w) = \max \left\{ |u_{ij} - w_{ij}|_v : 1 \leq i, j \leq 2 \right\}$ where the indices $i, j$ stand for the entries of the corresponding matrix. We usually denote the distance from $0 \in \mathfrak{g}_v$ by $\mathrm{d}_{\mathfrak{g}_v}(u, 0) = \|u\|$ and refer to it as the *norm* of $u$. We define the metrics $\mathrm{d}_{G_S}, \mathrm{d}_{\mathfrak{g}_S}$ on $G_S, \mathfrak{g}_S$ respectively by taking the maximum of the metrics defined above over the places in $S$. The metric $\mathrm{d}_{G_S}$ induces a right-$K_S$-invariant metric on $X_S$ by setting $\mathrm{d}_{X_S}(\Gamma_S g_1, \Gamma_S g_2) = \inf_{\gamma \in \Gamma_S} \mathrm{d}_{G_S}(\gamma g_1, g_2)$.

In a metric space $(X, \mathrm{d}_X)$ we denote $B_r^X(x)$ the open ball of radius $r$ around $x$. In case the space is a group, we denote by $B_r^X$ the corresponding ball around the trivial element.

We finish this section by stating two basic facts in the form of lemmas for convenience of reference.

**Lemma 3.2.** *Let $xH \subset X_S$ be a periodic orbit of a closed subgroup $H < G_S$ with Haar measure $\eta$, then for any $g \in G$ the translate $xHg = xgH^g$ is a periodic orbit for $H^g$ with Haar measure $g_* \eta$.*

**Lemma 3.3.** *Let $h(t)$ be a one parameter subgroup of $G_\infty$. Then for any $g \in G_\infty$, $\mathrm{d}_{G_\infty}(g, gh(t)) \leq \left\| \dot{h}(0) \right\| t$, where $\left\| \dot{h}(0) \right\|$ is the norm of the derivative of $h(t)$ at the identity.*

## 4. The $S$-Hecke graph and the main theorem

Throughout this section we use the notation introduced in §3 with the choice $\mathbb{G} = \mathrm{PGL}_2$. We fix a finite set of places $S$ containing $\infty$. The space $X_\infty$ can be thought of as the moduli space of equivalence classes of 2-dimensional lattices in the plane $\mathbb{R}^2$ up to homothety. We will refer below to a point $x \in X_\infty$ as a class; here, the class $\Gamma_\infty g$ is composed of the lattice spanned by the rows of the matrix $g$ (which is well defined up to scaling) and all its homotheties.

We begin by describing a setting which will put Theorem 2.8 in a context which will allow us to state a certain generalization of it. In a few words, we will fix a class $x$ with periodic $A_\infty$-orbit and consider a class $x'$ on the $S$-Hecke graph through $x$ and prove an effective equidistribution statement regarding the periodic orbit $x' A_\infty$ as $x'$ drifts away from $x$ in the graph.

4.1. **Hecke friends.** Given a class $x \in X_\infty$, we say that a class $x'$ is a *Hecke friend* of $x$ if one can choose lattices $\Lambda_x \in x, \Lambda_{x'} \in x'$ such that $\Lambda_{x'} < \Lambda_x$. After fixing the lattice $\Lambda_x$

there is a unique choice of $\Lambda_{x'} \in x'$ such that $\Lambda_{x'} < \Lambda_x$ is *primitive*; that is, such that the index $[\Lambda_x : \Lambda_{x'}]$ is minimal. We denote this minimal index by $\mathrm{ind}(x, x')$. We say that $x'$ is an *$S_f$-Hecke friend* of $x$ if the index $\mathrm{ind}(x, x')$ is supported on $S_f$. It is elementary to check that the Hecke friendship relation is an equivalence relation and that furthermore, if $x, x'$ are Hecke friends then $\mathrm{ind}(x, x') = \mathrm{ind}(x', x)$.

4.2. **The graph.** For a class $x \in X_\infty$ we define

$$\mathcal{G}_S(x) = \{x' \in X_\infty : x, x' \text{ are } S_f\text{-Hecke friends}\} \tag{4.1}$$

The set $\mathcal{G}_S(x)$ has the structure of a graph[6]: We join $x_1, x_2 \in \mathcal{G}_S(x)$ with an edge if there exists $\Lambda_i \in x_i$ such that $\Lambda_1$ is a sublattice of $\Lambda_2$ of index $p$ for some $p \in S_f$ (note that as $p$ is prime this forces $\Lambda_1$ to be a primitive sublattice of $\Lambda_2$). In this case we declare the length of this edge to be $\log(p)$. This induces a distance function on the graph which we denote $\mathrm{d}_{\mathcal{G}}(\cdot, \cdot)$ for which $\mathrm{d}_{\mathcal{G}}(x_1, x_2) = \log(\mathrm{ind}(x_1, x_2))$. We will refer to $x$ as the *root* of $\mathcal{G}_S(x)$ and call $\mathcal{G}_S(x)$ the *$S$-Hecke graph through $x$*. Note that the possible values of $\mathrm{ind}(x, x')$ are exactly the heights $\mathrm{ht}(q)$ where $q$ varies along the rationals supported on $S_f$ or in other words, the integers supported on $S_f$. We abuse the language often and refer to these heights as *admissible radii* and denote by $\mathcal{S}_{\mathrm{h}}(x) = \{x' \in \mathcal{G}_S(x) : \mathrm{ind}(x, x') = \mathrm{h}\}$ the *sphere* of radius h around the root $x$.

4.3. **The sphere.** For a rational $q$ supported on $S_f$ let us define

$$a_f(q) = \left(e_\infty, \begin{pmatrix} 1 & 0 \\ 0 & q \end{pmatrix}, \ldots, \begin{pmatrix} 1 & 0 \\ 0 & q \end{pmatrix}\right) \in G_S. \tag{4.2}$$

Given $x \in X_\infty$ we wish to have a convenient algebraic description of the classes on the sphere $\mathcal{S}_{\mathrm{h}}(x)$ for admissible radii h. We obtain this description using the extension $\pi : X_S \to X_\infty$ in the following way: Lemma 4.1 below shows that the various points on $\mathcal{S}_{\mathrm{h}}(x)$ are obtained by choosing a lift $y \in \pi^{-1}(x)$ of $x$, and projecting $y a_f(\mathrm{h})$ via $\pi$ back to $X_\infty$.

**Lemma 4.1.** *For $x = \Gamma_\infty g \in X_\infty$ and* h *an admissible radius we have*

$$\begin{aligned} \mathcal{S}_{\mathrm{h}}(x) &= \pi\left(\{\Gamma_S(g, \gamma) : \gamma \in \Gamma_\infty\} a_f(\mathrm{h})\right) \\ &= \pi\left(\pi^{-1}(x) a_f(\mathrm{h})\right). \end{aligned} \tag{4.3}$$

*Proof.* Recall that the elementary divisors theorem attaches to any pair of lattices $\Lambda_1 < \Lambda_2$ in the plane, a pair of integers $d_1, d_2$ which are characterized by the following two properties: (1) the divisibility $d_2 | d_1$ holds, (2) there exists a basis $v_1, v_2$ of $\Lambda_2$ such that $d_1 v_1, d_2 v_2$ forms a basis of $\Lambda_1$. Note that $\Lambda_1$ is a primitive sublattice of $\Lambda_2$ if and only if the second divisor satisfies $d_2 = 1$. We conclude from here that given a class $x = \Gamma_\infty g$, then a class $x'$ lies on the sphere $\mathcal{S}_{\mathrm{h}}(x)$ if and only if there exists a lattice $\Lambda_{x'} \in x'$ which

[6] When $S_f$ contains only one prime, this is the well known *$p$-Hecke tree* through $x$. In general, this graph is the product of the various $p$-Hecke trees for $p \in S_f$.

is a sublattice of the lattice $\Lambda_x \in x$ spanned by the rows of $g$ such that the elementary divisors are $d_1 = \mathrm{h}, d_2 = 1$. In other words we have the equality

$$\mathcal{S}_{\mathrm{h}}(x) = \left\{\Gamma_\infty \operatorname{diag}(\mathrm{h},1)\,\gamma g : \gamma \in \Gamma_\infty\right\}. \tag{4.4}$$

The following identity is crucial for us. It shows how the lattice $\Gamma_S$ causes the desired interaction between the real and $p$-adic components in the extension $X_S$ of $X_\infty$:

$$\begin{aligned}\Gamma_\infty \operatorname{diag}(\mathrm{h},1)\,\gamma g &= \pi\left(\Gamma_S(\operatorname{diag}(\mathrm{h},1)\,\gamma g, e_f)\right) \qquad (4.5)\\ &= \pi\left(\Gamma_S \underbrace{\gamma^{-1}\operatorname{diag}(1,\mathrm{h})}_{\in\Gamma_S}(\operatorname{diag}(\mathrm{h},1)\,\gamma g, e_f)\right) = \pi\left(\Gamma_S(g,\gamma^{-1})a_f(\mathrm{h})\right).\end{aligned}$$

From equations (4.4),(4.5) we immediately conclude that

$$\mathcal{S}_{\mathrm{h}} = \pi\left(\left\{\Gamma_S(g,\gamma) : \gamma \in \Gamma_\infty\right\} a_f(\mathrm{h})\right),$$

which is the first equality in (4.3). Using the first equality, the second equality follows once we show that for any given $\omega \in K_{S_f}$ there exist $\gamma \in \Gamma_\infty$ such that $\pi(\Gamma_S(g,\gamma)a_f(\mathrm{h})) = \pi(\Gamma_S(g,\omega)a_f(\mathrm{h}))$. A short calculation using Remark 3.1 shows that this happens precisely when

$$\gamma^{-1}\omega \in a_f(\mathrm{h})K_{S_f}a_f(\mathrm{h})^{-1}. \tag{4.6}$$

Thus, let $\omega = (\omega_p)_{p\in S_f} \in K_{S_f}$ be given and write $\omega_p = \theta_p \cdot \operatorname{diag}(1, \det(\omega_p))$, with $\theta_p \in \mathrm{SL}_2(\mathbb{Z}_p)$. Let $U_n^p < \mathrm{SL}_2(\mathbb{Z}_p)$ be the subgroup consisting of elements congruent to the identity modulo $p^n$. By the strong approximation Theorem for $\mathrm{SL}_2$ (see[PR94, §7.4]), for any $n \in \mathbb{N}$ there exist $\gamma_n \in \Gamma_\infty$ such that for all $p \in S_f$

$$\gamma_n^{-1}\theta_p \in U_n^p.$$

Note that there exist $N = N(\mathrm{h}) \in \mathbb{N}$ such that for all $n > N$ we have that the image of $\prod_{p\in S_f} U_n^p$ in $G_{S_f}$ lies in $a_f(\mathrm{h})K_{S_f}a_f(\mathrm{h})^{-1}$. As $a_f(\mathrm{h})K_{S_f}a_f(\mathrm{h})^{-1}$ is a group that contains $(\operatorname{diag}(1,\det(\omega_p)))_{p\in S_f}$ we conclude that $\prod_{p\in S_f} U_n^p \cdot \operatorname{diag}(1,\det(\omega_p)) \subset a_f(\mathrm{h})K_{S_f}a_f(\mathrm{h})^{-1}$. Therefore any $\gamma_n$ with $n > N$ will satisfy equation (4.6). This concludes the proof of the lemma.

□

**Definition 4.2.** Let $x \in X_\infty$ be given. Let $g_x \in G_\infty$ be a choice of a representative for $x$ so that $x = \Gamma_\infty g_x$. For any choice $\omega \in K_{S_f}$ we define the *generalized branch* $\mathcal{L}_{g_x,\omega} \subset \mathcal{G}_S(x)$ to be the set

$$\mathcal{L}_{g_x,\omega} = \pi\left(\left\{\Gamma_S(g_x,\omega)a_f(\mathrm{h}) : \mathrm{h} \text{ is an admissible radius}\right\}\right). \tag{4.7}$$

When $\omega$ is a rational element (i.e. for any $p \in S_f$ the $p$'th component $\omega_p$ of $\omega$ satisfies $\omega_p \in K_p \cap \mathrm{PGL}_2(\mathbb{Q})$) we call the generalized branch $\mathcal{L}_{g_x,\omega}$ a *rational generalized branch.*

The reader should think of the generalized branches as prescribed ways to go to infinity in the graph $\mathcal{G}_S(x)$. When $S_f$ is composed of a single prime the generalized branches are exactly the branches on the Hecke tree that start from the root $x$.

**Remark 4.3.** We wish point out a few things regarding the definition of generalized branches and fix some notation that will be used in the sequel. Let $x = \Gamma_\infty g_x \in X_\infty$ be given.

(1) For any $\omega \in K_{S_f}$ and any admissible radius h we denote $y_{\omega,\mathrm{h}} = \Gamma_S(g_x, \omega)a_f(\mathrm{h}) \in X_S$, $x_{\omega,\mathrm{h}} = \pi(y_{\omega,\mathrm{h}}) \in X_\infty$. The generalized branch $\mathcal{L}_{g_x,\omega}$ intersects the sphere $\mathcal{S}_\mathrm{h}(x)$ in a single point, namely
$$x_{\omega,\mathrm{h}} = \mathcal{L}_{g_x,\omega} \cap \mathcal{S}_\mathrm{h}(x). \tag{4.8}$$
When the generalized branch is fixed (that is when $\omega$ is fixed) we sometime denote $x_\mathrm{h} = x_{\omega,\mathrm{h}}$. We stress here the dependency on the representative $g_x$ of $x$. Note that we do not recall this dependency in the notation $x_{\omega,\mathrm{h}}, y_{\omega,\mathrm{h}}$.

(2) Two generalized branches $\mathcal{L}_{g_x,\omega_1}, \mathcal{L}_{g_x,\omega_2}$ intersect the sphere $\mathcal{S}_\mathrm{h}(x)$ at the same point, that is, $x_{\omega_1,\mathrm{h}} = x_{\omega_2,\mathrm{h}}$, if and only if the points $y_{\omega_i,\mathrm{h}}$ lie in the same fiber of $\pi$. This is in turn equivalent to saying that the conjugation $(\omega_2^{-1}\omega_1)^{a_f(\mathrm{h})}$ lies in $K_{S_f}$ (see Remark 3.1). This happens if and only if the lower left coordinate of each of the components of $\omega_2^{-1}\omega_1$ is divisible by h in the corresponding ring $\mathbb{Z}_p$. In particular, it follows that it is divisible by any integer that divides h which means by the same reasoning, that the two branches intersect all the spheres $\mathcal{S}_{\mathrm{h}'}$ at the same points, for any choice of admissible radius $\mathrm{h}'$ dividing h. Moreover, it follows from here that given $\omega_1, \omega_2 \in K_{S_f}$, the two generalized branches $\mathcal{L}_{g_x,\omega_i}$ are identical if and only if the quotient $\omega_2^{-1}\omega_1$ is an upper triangular element of $K_{S_f}$.

(3) From the above it follows that the collection of generalized branches may be identified with the quotient $K_{S_f}/B$, where $B < K_{S_f}$ denotes the group of upper triangular elements (this identification depends of course on the choice of the representative $g_x$).

(4) If we replace $g_x$ by another representative $\gamma g_x$ for $\gamma \in \Gamma_\infty$, then it readily follows that for any $\omega \in K_{S_f}$, $\mathcal{L}_{\gamma g_x,\omega} = \mathcal{L}_{g_x,\gamma^{-1}\omega}$. In particular, the notion of rationality of a generalized branch is well defined.

(5) Let $q$ be a rational supported on $S_f$ and write $q = \frac{\ell_1}{\ell_2}$ in reduced form (that is, $\ell_1, \ell_2$ are coprime). We let the *type* of $q$ be the subset $\tau_q \subset S_f$ defined by $\tau_q = \{p \in S_f : p|\ell_2\}$. Thus, the rationals supported on $S_f$ are partitioned into $2^{|S_f|}$ sets according to their types. We leave it as an exercise to the reader to show that for any class $x = \Gamma_\infty g \in X_\infty$ and any $\omega \in K_{S_f}$, the collection
$$\mathcal{L}^\tau_{g,\omega} = \pi\left(\{\Gamma_S(g,\omega)a_f(q) : q \text{ is a rational supported on } S_f \text{ of type } \tau\}\right)$$
is a single generalized branch and that the class $\pi(\Gamma_S(g,\omega)a_f(q))$ lies on the sphere $\mathcal{S}_{\mathrm{ht}(q)}(x)$. Moreover, the notion of rationality of the generalized branch is independent of the type; that is, $\mathcal{L}^\tau_{g,\omega}$ is rational if and only if $\omega$ may be chosen rational.

4.4. **Periodic $A_\infty$-orbits.** The following classical lemma relates the periodic $A_\infty$-orbits to quadratic irrationals.

**Lemma 4.4.** *Let $\alpha$ be a quadratic irrational, $g_\alpha \in G_\infty$ be as in* (2.5)*, and $x_\alpha = \Gamma_\infty g_\alpha \in X_\infty$. Then, the orbit $x_\alpha A_\infty \subset X_\infty$ is periodic.*

*Proof.* Consider the $\mathbb{Z}$-module $\Lambda_\alpha = \operatorname{span}_{\mathbb{Z}}\{1,\alpha\}$ in the field $\mathbb{Q}(\alpha)$. There exists a unit $\omega$ in the ring of integers which stabilizes $\Lambda_\alpha$ and furthermore by replacing $\omega$ by $\omega^2$ if necessary we may assume that both $\omega$ and its Galois conjugate $\omega'$ are positive. Note that the diagonal matrix $\operatorname{diag}(\omega,\omega')$ is an element of $A_\infty$. Let $\gamma = \begin{pmatrix} n & m \\ k & \ell \end{pmatrix} \in \mathrm{GL}_2(\mathbb{Z})$ be the matrix describing the passage from the basis $\{1,\alpha\}$ to the basis $\{\omega,\omega\alpha\}$ of $\Lambda_\alpha$. That is

$$\begin{pmatrix} n & m \\ k & \ell \end{pmatrix}\begin{pmatrix} \alpha \\ 1 \end{pmatrix} = \begin{pmatrix} \omega\alpha \\ \omega \end{pmatrix}. \tag{4.9}$$

The reader will easily verify now that (4.9) implies that $\gamma g_\alpha = g_\alpha \operatorname{diag}(\omega,\omega')$ or in other words that in the space $X_\infty$ the orbit $x_\alpha A_\infty$ is periodic as desired. □

**Remark 4.5.** In fact, it is well known (cf. [LW01],[McM05],[ELMV09]) that any periodic $A_\infty$-orbit in $X_\infty$ is of the above form. Given a periodic orbit $xA_\infty$, we denote by $\mathbb{F}_x$ the corresponding quadratic extension of $\mathbb{Q}$ from which the periodic orbit $xA_\infty$ arises (see also Remark 7.4 below).

**Definition 4.6.** Given $x \in X_\infty$ with a periodic $A_\infty$-orbit we denote by $t_x$ the *length of the period*, i.e. the minimal positive $t$ for which $xa_\infty(t) = x$. If $x = x_\alpha$ (in the notation introduced after (2.5)), we denote this period by $t_\alpha$. We let $\mu_x$ denote the unique $A_\infty$-invariant probability measure supported on $xA_\infty$ and similarly, when $x = x_\alpha$ we denote this measure by $\mu_\alpha$.

Let $x \in X_\infty$ be a class with a periodic $A_\infty$-orbit. It is straightforward to argue that any $x' \in \mathcal{G}_S(x)$ has a periodic orbit as well. We are interested in understanding the way the orbit $x'A_\infty$ is distributed in $X_\infty$ as $\mathrm{d}_{\mathcal{G}}(x,x')$ goes to $\infty$.

**Remark 4.7.** It turns out that the answer to this question has to do with the question of whether or not the primes $p \in S_f$ split in the field $\mathbb{F}_x$ (see Remark 4.5). Let $\gamma \in \Gamma_\infty$ be a matrix such that the roots of its characteristic polynomial generate $\mathbb{F}_x$. Recall that a prime $p$ splits in $\mathbb{F}_x$ if and only if $\gamma$ is diagonalizable over $\mathbb{Q}_p$. A short exercise in linear algebra shows that $\gamma \in \Gamma_\infty$ is diagonalizable over $\mathbb{Q}_p$ if and only if it can be triangulized over $\mathbb{Z}_p$.

**Definition 4.8.** Let $x$ be a class with a periodic $A_\infty$-orbit and fix a representative $g_x$ so that $x = \Gamma_\infty g_x$. Let $\gamma_x \in \Gamma_\infty$ be the matrix satisfying $\gamma_x g_x = g_x a_\infty(t_x)$, and let $\omega \in K_{S_f}$.

(1) We say that the generalized branch $\mathcal{L}_{g_x,\omega}$ is *degenerate* (for $S_f$) if there exists $p \in S_f$ such that $\omega_p^{-1}\gamma_x^n\omega_p$ is upper triangular for some positive integer $n$ (here $\omega_p$ is the $p$'th component of $\omega \in K_{S_f}$).[7]

(2) We say that the class $x$ is *split* (for $S_f$) if there exists $p \in S_f$ which splits over $\mathbb{F}_x$ (or equivalently, by Remark 4.7, if there exists a degenerate generalized branch).

We are now ready to state the main result of the present paper.

[7] This is equivalent to saying that the Lie algebra of the closure of the group generated by $\gamma_x^\omega$ in $G_p$ is upper triangular.

**Theorem 4.9.** *Let $x = \Gamma_\infty g_x \in X_\infty$ be such that $xA_\infty$ is periodic.*

1. *Let $\mathcal{L} = \mathcal{L}_{g_x,\omega}$ be a nondegenerate generalized branch of the graph $\mathcal{G}_S(x)$ and $\mathrm{h}$ an admissible radius, then for any $\kappa$-Lipschitz function $\varphi_0 \in L^2(X_\infty, m_\infty)$ and any $\epsilon > 0$ the following holds*
$$\left| \int_{X_\infty} \varphi_0 d\mu_{x_{\mathrm{h}}} - \int_{X_\infty} \varphi_0 dm_\infty \right| \ll_{x,S_f,\mathcal{L},\epsilon} \max\left\{ \left\|\varphi_0\right\|_2, \kappa \right\} \mathrm{h}^{-\frac{\delta_0}{2}+\epsilon}. \tag{4.10}$$
2. *If $x$ is non-split (i.e. all generalized branches are non-degenerate), the implicit constant in* (4.10) *may be chosen to be independent of the generalized branch and we have uniform rate of equidistribution along the full graph.*
3. *If $x$ is split and $\mathcal{L}_{g_x,\omega}$ is a degenerate generalized branch, then there is a sequence of admissible radii $\mathrm{h}_n \to \infty$ such that for the sequence of classes $x_{\mathrm{h}_n}$, the lengths $t_{x_{\mathrm{h}_n}}$ of the orbits $x_{\mathrm{h}_n} A_\infty$ are bounded and in particular, the orbits do not equidistribute.*
4. *Rational generalized branches are always nondegenerate and so* (4.10) *holds automatically.*
5. *Nonetheless, in case $x$ is split, the implicit constants in* (4.10) *cannot be taken to be uniform for the rational generalized branches.*

**Remark 4.10.** Let $\alpha$ be a quadratic irrational and $q$ a rational supported on $S_f$. Suppose for example that $\alpha > \alpha'$ so that the matrix $g_\alpha$ has the form that appears on the left hand side of (2.5). We wish to analyze where on $\mathcal{G}_S(x_\alpha)$ the class $x_{q\alpha}$ lies. We have the following identity

$$\begin{aligned} x_{q\alpha} = \pi\left(\Gamma_S(g_{q\alpha}, e_f)\right) = & \qquad\qquad\qquad\qquad\qquad\qquad\qquad (4.11) \\ & \pi\left(\Gamma_S \underbrace{a_\infty(q^{-1})a_f(q^{-1})}_{\in\Gamma_S}(g_\alpha, e_f)a_f(q)\right) = \pi\left(\Gamma_S(g_\alpha, e_f)a_f(q)\right), \end{aligned}$$

which shows that $x_{q\alpha}$ lies on the rational generalized branch $\mathcal{L}^{\tau_q}_{g_\alpha,e_f}$ on the sphere $\mathcal{S}_{\mathrm{ht}(q)}(x_\alpha)$ (see Remark 4.3(5)). As only finitely many such generalized branches are involved we see that parts (1),(4) of Theorem 4.9 indeed imply Theorem 2.8.

4.5. **Walking on the tree.** Let us consider for simplicity the case when $S_f$ consists of a single prime $p$ and let $x = \Gamma_\infty g_x$ be a class in $X_\infty$ with a fixed representative $g_x$. As noted earlier, the graph $\mathcal{G}_p(x)$ is a $p+1$-regular tree. Let us denote

$$S_p = \begin{pmatrix} p & 0 \\ 0 & 1 \end{pmatrix},\ T_{p,j} = \begin{pmatrix} 1 & j \\ 0 & p \end{pmatrix}, j = 0, 1 \ldots p-1. \tag{4.12}$$

Furthermore, we set $\mathbf{H} = \{S_p, T_{p,0}, \ldots, T_{p,p-1}\}$.

**Definition 4.11.** Let us refer to a sequence $R_1, \ldots R_n$, $R_i \in \mathbf{H}$ as *legitimate* if any appearance of the element $S_p$ does not follow the appearance of any element of the form $T_{p,j}$, and, any appearance of the element $T_{p,0}$ does not follow an appearance of $S_p$.

A quick computation shows that the number of legitimate sequences of length $n$ is $(p+1)p^{n-1}$ which is exactly the cardinality of the sphere $\mathcal{S}_{p^n}(x)$. The following lemma shows that this is not a coincidence. We leave the proof as an exercise to the reader.

**Lemma 4.12.** *There is an equality*

$$\mathcal{S}_{p^n}(x) = \{\Gamma_\infty R_n \ldots R_1 g_x : R_1, \ldots, R_n \text{ is legitimate}\}.$$

*Furthermore, the infinite sequences $\{R_i\}$ of the elements of $\boldsymbol{H}$ describe all possible paths on the graph $\mathcal{G}_p(x)$ that start from the root $x$ and a sequence is legitimate if and only if the path it describes is a branch. Finally, if $\{R_i\}$ is a legitimate sequence which is eventually periodic, then it corresponds to a rational branch.*

Let $\alpha$ be a quadratic irrational and $x_\alpha = \Gamma_\infty g_\alpha$ as before. Given a legitimate sequence $\{R_n\}_{n=0}^\infty$, $R_n \in \mathbf{H}$, we consider the resulting recursive sequence of numbers

$$\alpha_{-1} = \alpha, \;\; \alpha_n = R_n \cdot \alpha_{n-1} \, n \geq 0, \tag{4.13}$$

where here $R_n$ acts on the real line as a Möbius transformation. Clearly $\alpha_n \in \mathbb{Q}(\alpha)$. We alternate between thinking of the elements $R_n$ of the sequence as describing a walk along a branch of the graph $\mathcal{G}_p(x_\alpha)$ and describing a sequence of arithmetic operations which result in the sequence $\alpha_n$ from (4.13). We now explain the connection between these two viewpoints. Given $h \in \mathrm{PGL}_2(\mathbb{Q})$ the reader should check that $hg_\alpha$ is obtained from $g_{h\cdot\alpha}$ by multiplication from the right by a diagonal element. In our case, $h$ is one of the elements of $\mathbf{H}$ and the calculation shows that this diagonal element lies in $A_\infty$. The following lemma follows

**Lemma 4.13.** *Given a legitimate sequence $\{R_n\}$ corresponding to the branch $\mathcal{L}_{g_\alpha,\omega} \subset \mathcal{G}_p(x_\alpha)$, let $\alpha_n$ be the resulting sequence of quadratic irrationals given in* (4.13)*, then $x_{\alpha_n} A_\infty = x_{\omega,p^n} A_\infty$.*

**Remark 4.14.** Lemmas 4.12,4.13, when combined with Theorem 4.9, imply the effective equidistribution of the periodic orbits $x_{\alpha_n} A_\infty$ for more general sequences than just $p^{\pm n}\alpha$ that were considered so far. For example, given any sequence $\{j_n\}_{n=0}^\infty$, where $j_n \in \{0, \ldots, p-1\}$, a quick calculation shows that the sequence of quadratic irrationals $\alpha_n$ corresponding to the legitimate sequence of operations given by $\{T_{p,j_n}\}$, is given by

$$\alpha_n = \frac{\alpha + \sum_{i=0}^{n-1} j_i p^i}{p^n}.$$

If $p$ does not split in $\mathbb{Q}(\alpha)$, then Theorem 4.9 (combined with Lemmas 4.12,4.13) imply the equidistribution $\mu_{\alpha_n} \to m_\infty$ (in the effective manner given in (4.10)). If $p$ splits in $\mathbb{Q}(\alpha)$, a more restrictive result holds; namely, in order to ensure that the corresponding branch is nondegenerate we confine ourselves to sequences $\alpha_n$ arising from eventually periodic sequences $j_n$.

## 5. Relations to other arguments

Before turning to the proof of Theorem 4.9 we wish to make some comments that will clarify its relation to arguments giving equidistribution of collections of periodic orbits. The result of Benoist and Oh [BO07, Theorem 1.1] imply that given a class $x$ with a periodic $A_\infty$-orbit, then the collection of orbits $\{x'A_\infty : x' \in \mathcal{S}_{\mathrm{h}}(x)\}$ (counted without multiplicities) is becoming equidistributed as $\mathrm{h} \to \infty$.

Ignoring the effectivity of Theorem 4.9 and just interpreting it as saying that $\mu_{x'} \to m_\infty$ as $x'$ drifts away from the root $x$ along a nondegenerate generalized branch, it seems tempting to think that it is considerably stronger than the result of Benoist and Oh, as it deals with the equidistribution of single orbits as opposed to the equidistribution of the full collection. We will show in §5.1 below that this 'soft' equidistribution in fact follows quite elementarily from the work of Benoist and Oh. Nonetheless, as far as we know the effective statement does not follow easily from known results.

5.1. **Total vs. individual growth.** Let $x \in X_\infty$ be a class with a periodic $A_\infty$-orbit and consider the union of the periodic orbits $x'A_\infty$ for $x' \in \mathcal{S}_{\mathrm{h}}(x)$ (where h is an admissible radius). We denote the total length of this union by $\mathbf{t}_x(\mathrm{h})$; that is, $\mathbf{t}_x(\mathrm{h}) = \sum t_{x'}$ where the sum is taken over a set of representatives of the classes on the sphere giving rise to different orbits. The following lemma shows that the growth rate of the length of individual periodic orbits along a nondegenerate generalized branch is the same as the growth rate of the total length. To some extent this phenomena is what stands behind our results. We do not really need this lemma but we state it here because we think it explains the phenomena at hand in a clear way. Nevertheless, in the course of deducing Corollary 2.5, we will need a part of it and for completeness, we provide the proof in §7.

**Lemma 5.1.** *Let $x \in X_\infty$ be a class with a periodic $A_\infty$-orbit and $\mathcal{L}$ a nondegenerate generalized branch of $\mathcal{G}_S(x)$.*

1. *The total length $\boldsymbol{t}_x(\mathrm{h})$ satisfies $\mathrm{h} \ll_{x,S_f} \boldsymbol{t}_x(\mathrm{h}) \ll_{x,S_f} \mathrm{h}$.*
2. *Let $c(\mathrm{h}) = c_{\mathcal{L}}(\mathrm{h})$ be the function defined by the equation $t_{x_{\mathrm{h}}} = c(\mathrm{h})\,\mathrm{h}$, where $x_{\mathrm{h}}$ is the class in $\mathcal{S}_{\mathrm{h}} \cap \mathcal{L}$. Then, $c(\mathrm{h})$ attains only finitely many values and moreover, if $\mathrm{h}_n$ is a divisibility sequence of admissible radii (that is $\mathrm{h}_n \mid \mathrm{h}_{n+1}$), then $c(\mathrm{h}_n)$ stabilizes.*
3. *The class $x$ is non-split for $S_f$ if and only if*

$$\inf\{c_{\mathcal{L}}(\mathrm{h}) : \mathcal{L} \textit{ nondegenerate}, \mathrm{h} \textit{ admissible radii}\} > 0. \tag{5.1}$$

The first two parts of Lemma 5.1 show that if $\mathcal{L}$ is a nondegenerate generalized branch, then a single orbit $x_{\mathrm{h}}A_\infty$ through the class $x_{\mathrm{h}} \in \mathcal{L} \cap \mathcal{S}_{\mathrm{h}}(x)$ actually occupies a positive proportion (bounded below by a constant independent of h) of the full collection $\{x'A_\infty : x' \in \mathcal{S}_{\mathrm{h}}(x)\}$. Relying on [BO07] we may argue the soft version of Theorem 4.9 (that is, that $\mu_{x_{\mathrm{h}}} \to m_\infty$ as $\mathrm{h} \to \infty$) in the following way: Let $\mathrm{h}_i \to \infty$ be a sequence of admissible radii such that $\mu_{x_{\mathrm{h}_i}}$ converges to say $\mu_\infty$ (which is an $A_\infty$-invariant measure). We need to argue that $\mu_\infty = m_\infty$. Let $\eta_{\mathrm{h}}$ be the natural $A_\infty$-invariant probability measure supported on the collection of periodic orbits $\{x'A_\infty : x' \in \mathcal{S}_{\mathrm{h}}(x)\}$. By [BO07]

$\eta_{\mathrm{h}} \to m_\infty$. By the first two parts of Lemma 5.1 we can write $\eta_{\mathrm{h}_i}$ as a convex combination of $A_\infty$-invariant probability measures in the following way: $\eta_{\mathrm{h}_i} = c'_{\mathrm{h}_i}\mu_{x_{\mathrm{h}_i}} + (1 - c'_{\mathrm{h}_i})\nu_{\mathrm{h}_i}$, where the constants $c'_{\mathrm{h}_i}$ are bounded below by some constant $c'$ independent of $\mathrm{h}_i$. Taking $i$ to $\infty$ (along an appropriate subsequences if necessary) we deduce that in the limit $m_\infty = c'_\infty\mu_\infty + (1 - c'_\infty)\nu_\infty$ for some positive constant $c'_\infty \leq 1$. By the ergodicity of $m_\infty$ with respect to the $A_\infty$-action we deduce that the limit $\mu_\infty$, that appears in the above convex combination with positive weight, must be equal to $m_\infty$. This establishes the desired convergence. We note here, as mentioned above, that even if one starts with an effective statement regarding the equidistribution of the measures $\mu_{\mathrm{h}}$ supported on the collection of orbits, it is not clear how to use the above argument to deduce an effective statement such as in Theorem 4.9 regarding the equidistribution of a positive fraction of the collection.

**Remark 5.2.** As noted in the introduction, in the above argument, instead of appealing to [BO07], one can appeal to Duke's Theorem [Duk88],[ELMV].

## 6. Proof of Theorem 4.9.

Throughout this section we fix $x \in X_\infty$ to be a class with a periodic $A_\infty$-orbit and a representative $g_x \in G_\infty$ such that $x = \Gamma_\infty g_x$. Using the notation of Definition 4.6, it follows that there exists $\gamma_x \in \Gamma_\infty$ such that

$$\gamma_x g_x a_\infty(t_x) = g_x. \tag{6.1}$$

We briefly discuss the relations between the various parts of Theorem 4.9. As the eigenvectors of $\gamma_x$ are irrational (and not roots of unity) it follows that $\gamma_x$ (or any of its powers) is not triangulizable over $\mathbb{Q}$ and so all the rational generalized branches are nondegenerate. Therefore part (4) of the theorem follows from part (1). Part (5) of the theorem follows from part (3) because of (4.3) which shows that any class on the $S$-Hecke graph $\mathcal{G}_S(x)$ lies on a rational generalized branch; the sequence $x_{\mathrm{h}_n}$ produced by part (3) may be viewed as a sequence of classes lying on (varying) rational generalized branches, showing that a uniform implicit constant for all rational generalized branches in (4.10) is impossible.

We begin with the necessary preparations for the arguments yielding parts (1),(2), and (3). We will see below that part (3) is a simple observation once the stage is set correctly and so the main bulk of the theorem lies in establishing parts (1) and (2).

After fixing $g_x$ we fix a generalized branch in $\mathcal{G}_S(x)$; that is, we fix an element $\omega \in K_{S_f}$ and set $\mathcal{L}_\omega = \mathcal{L}_{g_x,\omega}$. Although $\omega$ is fixed, the reader should bear in mind that at some point we will vary the choice of $\omega$ in order to change the generalized branch.

6.1. **The lift of a closed loop.** The following construction is fundamental to our argument. Let $y_\omega \in X_S$ be defined by $y_\omega = \Gamma_S(g_x, \omega)$. Consider the orbit $y_\omega A_\infty \subset X_S$ and note that

$$\pi(y_\omega) = x,\ xA_\infty = \pi(y_\omega A_\infty) = \pi(\overline{y_\omega A_\infty}), \tag{6.2}$$

where the rightmost equality follows from the fact that $xA_\infty$ is compact and the continuity of the projection $\pi$.

We now analyze the closure $\overline{y_\omega A_\infty}$. Each $t \in \mathbb{R}$ can be written in a unique way in the form $t = s + \ell t_x$ for some $s \in [0, t_x)$ and $\ell \in \mathbb{Z}$. It follows from (6.1) that

$$y_\omega a_\infty(t) = \Gamma_S(g_x a_\infty^\ell(t_x) a_\infty(s), \omega) = \Gamma_S(g_x a_\infty(s), \gamma_x^\ell \omega) = y_\omega(a_\infty(s), \omega^{-1}\gamma_x^\ell \omega). \tag{6.3}$$

If we denote for an element $\gamma$ in a group $H$ by $\langle\gamma\rangle_H$ the cyclic group generated by $\gamma$ in $H$, then it follows from (6.3) that

$$y_\omega A_\infty = y_\omega(A_\infty \times \langle \omega^{-1}\gamma_x\omega\rangle_{G_{S_f}}). \tag{6.4}$$

Let

$$H_\omega = \omega^{-1}\overline{\langle\gamma_x\rangle}_{G_{S_f}}\omega = \overline{\langle\omega^{-1}\gamma_x\omega\rangle}_{G_{S_f}}. \tag{6.5}$$

Clearly, $H_\omega$ is a compact subgroup of $K_{S_f}$. We let

$$L_\omega = A_\infty \times H_\omega. \tag{6.6}$$

**Lemma 6.1.** *The orbit $y_\omega L_\omega$ is compact and*

$$\overline{y_\omega A_\infty} = y_\omega L_\omega. \tag{6.7}$$

*Proof.* We first establish (6.7). The inclusion $\supset$ follows readily from (6.4). For the reverse inclusion, let $t_n \in \mathbb{R}$ be such that $y_\omega a_\infty(t_n) \to_{n\to\infty} y \in \overline{y_\omega A_\infty}$. Let $s_n \in [0, t_x), \ell_n \in \mathbb{Z}$ be as defined before (6.3); that is $t_n = s_n + \ell_n$. By compactness we may assume without loss of generality (after passing to a subsequence if necessary) that $s_n \to s$ and $\omega^{-1}\gamma_x^{\ell_n}\omega \to h$. We conclude from (6.3) that

$$y = \lim y_\omega a_\infty(t_n) = \lim y_\omega(a_\infty(s_n), \omega^{-1}\gamma_x^{\ell_n}\omega) = y_\omega(a_\infty(s), h) \in y_\omega L_\omega. \tag{6.8}$$

The fact that the orbit $y_\omega L_\omega$ is compact now follows from the fact that it is a closed set contained in $\pi^{-1}(xA_\infty)$ which is compact by the properness of $\pi$. □

**Remark 6.2.** The above proof actually establishes a bit more: We have shown that in fact,

$$\overline{y_\omega A_\infty} = y_\omega L_\omega = \{y_\omega(a_\infty(t), h) : t \in [0, t_x), h \in H_\omega\}. \tag{6.9}$$

**Definition 6.3.** Let $\eta_\omega$ denote the $L_\omega$-invariant probability measure supported on the compact (and hence periodic) orbit $y_\omega L_\omega$. For an admissible radius $\mathrm{h}$ let

$$y_{\omega,\mathrm{h}} = y_\omega a_f(\mathrm{h}),\ L_\omega^{a_f(\mathrm{h})} = L_{\omega,\mathrm{h}},\ H_{\omega,\mathrm{h}} = H_\omega^{a_f(\mathrm{h})},$$

and note the identity $y_\omega L_\omega a_f(\mathrm{h}) = y_{\omega,\mathrm{h}} L_{\omega,\mathrm{h}} = y_{\omega,\mathrm{h}}(A_\infty \times H_{\omega,\mathrm{h}})$. We denote the unique $L_{\omega,\mathrm{h}}$-invariant probability measure supported on the periodic orbit $y_{\omega,\mathrm{h}} L_{\omega,\mathrm{h}}$ by $\eta_{\omega,\mathrm{h}}$. By Lemma 3.2 we have that $(a_f(\mathrm{h}))_*\eta_\omega = \eta_{\omega,\mathrm{h}}$. Note that the notation $y_{\omega,\mathrm{h}}$ is consistent with the one introduced in Remark 4.3(1).

**Lemma 6.4.** *Let $\mathrm{h}$ be an admissible radius and $x_{\omega,\mathrm{h}} \in \mathcal{L}_\omega \cap \mathcal{S}_\mathrm{h}(x)$. Then, the pushed orbit $y_\omega L_\omega a_f(\mathrm{h}) = y_{\omega,\mathrm{h}} L_{\omega,\mathrm{h}}$ projects to the periodic orbit $x_{\omega,\mathrm{h}} A_\infty$ and furthermore, the measure $\eta_{\omega,\mathrm{h}}$ supported on it is a lift of $\mu_{x_{\omega,\mathrm{h}}}$; i.e. $\pi_*\eta_{\omega,\mathrm{h}} = \mu_{x_{\omega,\mathrm{h}}}$.*

The above lemma puts us in a desirable situation from the dynamical point of view; instead of studying the orbits $x'A_\infty$ in the space $X_\infty$ as $x'$ drifts away from $x$ on a generalized branch (the connection between which is not clear apriori), we will study the images of the fixed orbit $y_\omega L_\omega$ under the action of $a_f(\mathrm{h})$ for admissible radii h, which share a clear algebraic (and geometric) relation. This relation is the reason we needed to introduce the $S$-arithmetic extension $X_S$.

*Proof.* The fact that $x_{\omega,\mathrm{h}} = \pi(y_{\omega,\mathrm{h}})$ follows from Definition 6.3 and Remark 4.3(1). We have that

$$\begin{aligned}\pi(y_{\omega,\mathrm{h}}L_{\omega,\mathrm{h}}) &= \pi(y_\omega L_\omega a_f(\mathrm{h})) = \pi(\overline{y_\omega A_\infty a_f(\mathrm{h})}) \\ &= \overline{\pi(y_\omega a_f(\mathrm{h})A_\infty)} = \overline{\pi(y_{\omega,\mathrm{h}})A_\infty} = x_{\omega,\mathrm{h}}A_\infty,\end{aligned} \tag{6.10}$$

where first equality from the left follows from Definition 6.3, the second, from Lemma 6.1 and that fact that $a_f(\mathrm{h})$ acts on $X_S$ by a homeomorphism, the third, from the commutation of $A_\infty$ and $a_f(\mathrm{h})$ and from the continuity of $\pi$, the fourth, from the fact that $\pi$ intertwines the $A_\infty$-actions on $X_S, X_\infty$, and finally the fifth equality follows from the fact that the orbit $x_{\omega,\mathrm{h}}A_\infty$ is compact.

As $A_\infty < L_{\omega,\mathrm{h}}$, $\eta_{\omega,\mathrm{h}}$ is $A_\infty$-invariant. As a consequence, the projection $\pi_*\eta_{\omega,\mathrm{h}}$ is an $A_\infty$-invariant probability measure supported on $x_{\omega,\mathrm{h}}A_\infty$. As $\mu_{x_{\omega,\mathrm{h}}}$ is the unique such measure, we conclude that $\pi_*\eta_{\omega,\mathrm{h}} = \mu_{x_{\omega,\mathrm{h}}}$ as desired. □

**Remark 6.5.** It follows from (6.9) and the definition of $y_{\omega,\mathrm{h}}, H_{\omega,\mathrm{h}}$ that

$$y_{\omega,\mathrm{h}}L_{\omega,\mathrm{h}} = \{y_{\omega,\mathrm{h}}(a_\infty(t), h) : t \in [0, t_x), h \in H_{\omega,\mathrm{h}}\}.$$

By (6.10) the following equality follows:

$$x_{\omega,\mathrm{h}}A_\infty = \pi\left(\{y_{\omega,\mathrm{h}}(a_\infty(t), h) : t \in [0, t_x), h \in H_{\omega,\mathrm{h}}\}\right). \tag{6.11}$$

The meaning of the above equation is that the only reason for the orbit $x_{\omega,\mathrm{h}}A_\infty$ to become long is that the group $H_{\omega,\mathrm{h}}$ stretches and 'sticks out' of $K_{S_f}$. This is illustrated in the following proof.

*Proof of part* (3) *of Theorem 4.9.* For an admissible radius h and $p \in S_f$ denote by $(H_{\omega,\mathrm{h}})_p$ the projection of the group $H_{\omega,\mathrm{h}}$ on its $p$-th component. Note that by definition, $(H_{\omega,\mathrm{h}})_p = \mathrm{diag}\left(1, \mathrm{h}^{-1}\right)(H_\omega)_p\,\mathrm{diag}\,(1, \mathrm{h})$.

Assume that the generalized branch $\mathcal{L}_\omega$ is degenerate. It follows that there exists $p \in S_f$ for which some power of the $p$-th component $(\omega^{-1}\gamma_x\omega)_p$ is upper triangular. Let $d$ be the minimal positive integer for which $(\omega^{-1}\gamma_x^d\omega)_p$ is upper triangular. We conclude from (6.5) that $(H_\omega)_p$ contains an index $d$ subgroup that consists of upper triangular elements only. Choose $\mathrm{h}_n = p^n$ and note that because of the above $(H_{\omega,\mathrm{h}_n})_p \cap K_p$ is of index at most $d$ in $(H_{\omega,\mathrm{h}_n})_p$. Moreover, note that as $p$ is a unit in $\mathbb{Z}_{p'}$ for any prime $p' \neq p$, we have that $(H_{\omega,\mathrm{h}_n})_{p'} < K_{p'}$. It follows that along the chosen sequence $\mathrm{h}_n$ we have that $H_{\omega,\mathrm{h}_n} \cap K_{S_f}$ has at most index $d$ in $H_{\omega,\mathrm{h}_n}$. Let $h_i \in H_{\omega,\mathrm{h}_n}, i = 1 \ldots d'$, $d' \leq d$, be representatives of the cosets of $H_{\omega,\mathrm{h}_n} \cap K_{S_f}$ and denote $y_i = y_{\omega,\mathrm{h}_n}h_i$, $i = 1 \ldots d'$ and $x_i = \pi(y_i)$. We can rewrite (6.11) as

$$\begin{aligned} x_{\omega,\mathrm{h}_n}A_\infty &= \pi\left(\cup_{i=1}^{d'}\left\{y_{\omega,\mathrm{h}_n}(a_\infty(t),h_ih) : t\in[0,t_x), h\in H_{\omega,\mathrm{h}_n}\cap K_{S_f}\right\}\right)\\ &= \pi\left(\cup_{i=1}^{d'}\left\{y_i(a_\infty(t),h) : t\in[0,t_x), h\in H_{\omega,\mathrm{h}_n}\cap K_{S_f}\right\}\right)\\ &= \cup_{i=1}^{d'}\left\{x_ia_\infty(t) : t\in[0,t_x)\right\}, \end{aligned} \tag{6.12}$$

and so we conclude that $t_{x_{\omega,\mathrm{h}_n}} \leq d't_x$ which finishes the proof. □

In order to finish the proof of Theorem 4.9 we are left to argue parts (1),(2). As said before, these are the main parts of the theorem.

6.2. **Strategy of the proof of Theorem 4.9**(1),(2)**.** In the notation of Lemma 6.4, because $\pi_*\eta_{\omega,\mathrm{h}} = \mu_{x_{\omega,\mathrm{h}}}$, the validity of (4.10) is equivalent to saying that given $\varphi \in L^2(X_S,m_S)$ which is $K_{S_f}$-invariant (i.e. is of the form $\varphi_0\circ\pi$ for $\varphi_0\in L^2(X_\infty,m_\infty)$) and $\kappa$-Lipschitz, then

$$\left|\int\varphi d\eta_{\omega,\mathrm{h}} - \int\varphi dm_S\right| \ll_{x,S_f,\mathcal{L}_\omega,\epsilon} \max\left\{\kappa,\|\varphi\|_2\right\}\mathrm{h}^{-\frac{\delta_0}{2}+\epsilon}. \tag{6.13}$$

The argument giving this 'effective equidistribution' is a combination of an argument which we will refer to as *the mixing trick* and spectral gap (or effective decay of matrix coefficients). As far as we know the mixing trick originates from Margulis' thesis [Mar04]. We briefly describe its heuristics: One slightly thickens the initial orbit $y_\omega L_\omega$ to an open set $\mathcal{T}\subset X_S$ in directions which are (weakly) contracted by the action of $a_f(\mathrm{h})$. The set $\mathcal{T}$ will be called below *a tube around the orbit* $y_\omega L_\omega$. Let $m_\mathcal{T}$ denote the normalized restriction of $m_S$ to $\mathcal{T}$. The pushed measure $(a_f(\mathrm{h}))_*m_\mathcal{T}$ is the normalized restriction of $m_S$ to the pushed tube $\mathcal{T}a_f(\mathrm{h})$, which is a tube around the orbit $y_{\omega,\mathrm{h}}L_{\omega,\mathrm{h}}$. Because the thickening used to construct $\mathcal{T}$ is taken in directions which are (weakly) contracted by $a_f(\mathrm{h})$, the size of the thickening giving the tube $\mathcal{T}a_f(\mathrm{h})$ is even smaller than the size of the initial thickening. Hence, there shouldn't be much of a difference between integrating against the measure $\eta_{\omega,\mathrm{h}}$ and integrating against $(a_f(\mathrm{h}))_*m_\mathcal{T}$. The fact that the action of $a_f(\mathrm{h})$ is mixing on $(X_S,m_S)$ means that the pushed measure $(a_f(\mathrm{h}))_*m_\mathcal{T}$ is 'close' to $m_S$ (here, the effective mixing Theorem 6.6 will allow us to pin down the meaning of 'close' in a precise way). Combining these things together will give us the desired estimate given in (6.13).

In order to make this strategy into a rigorous proof we discuss in the next two subsections in detail the construction of tubes and decay of matrix coefficients.

6.3. **Effective mixing.** Let $\mathcal{H} = L^2(X_S,m_S)$. Our goal in this section is to prove the following:

**Theorem 6.6.** *Let* $\mathrm{h}$ *be an admissible radius and* $w_1,w_2\in\mathcal{H}$ *be vectors with the following properties:* $w_1$ *is* $K_{S_f}$*-fixed and* $w_2$ *is stabilized by a product subgroup* $K^* = \prod_{v\in S_f}K_v^* <$

*$K_{S_f}$ of index $d$ in $K_{S_f}$. Then for any $\epsilon > 0$,*

$$|\langle w_1, a_f(\mathrm{h})w_2\rangle - \langle w_1, 1\rangle\langle 1, w_2\rangle| \ll_\epsilon \|w_1\| \, \|w_2\| \, d^{\frac{1}{2}} \, \mathrm{h}^{-\delta_0+\epsilon}, \tag{6.14}$$

The meaning of the exponent $\delta_0$ that appears in (6.14) will be explicated shortly. Before turning to the proof of the above theorem, we need to discuss three lemmas. For $v \in S$ let $\mathcal{H}_v$ denote the orthocomplement of the $G_v$-invariant functions in $\mathcal{H}$. The following is [Ven10, Lemma 9.1]. It is the key input in the proof of Theorem 6.6.

**Lemma 6.7.** *Let $w_1, w_2 \in \mathcal{H}_v$ ($v \in S_f$) be two vectors which are stabilized respectively by finite index subgroups $K^{(1)}, K^{(2)}$ of $K_v$, let $d_i = [K_v, K^{(i)}]$, and $a_v(t) = \operatorname{diag}(1, t), t \in \mathbb{Q}_v^*$. Then the following holds*

$$|\langle w_1, a_v(t)w_2\rangle| \ll_\epsilon \|w_1\| \, \|w_2\| \, d_1^{\frac{1}{2}} d_2^{\frac{1}{2}} \max\left\{|t|_v, \left|t^{-1}\right|_v\right\}^{-\delta_0+\epsilon}. \tag{6.15}$$

The exponent $\delta_0$ comes from the following discussion. Let $\rho_v$ be the unitary representation of $G_v$ on $\mathcal{H}_v$. Let $\sigma_0$ be the smallest number so that no complementary series representations of parameter $\geq \sigma_0$ is weakly contained in $\rho_v$. Here we follow [Ven10] and parametrize the complementary series representations by the parameter $\sigma \in (0, \frac{1}{2})$; so $\sigma_0 = 0$ corresponds to $\rho_v$ being tempered (the Ramanujan conjecture) and $\sigma_0 = \frac{1}{2}$ corresponds to $\rho_v$ having no almost invariant vectors. The best bound known today towards Ramanujan is given by Kim and Sarnak in the appendix of [Kim03] and establishes the bound $\sigma_0 \leq \frac{7}{64}$. The exponent $\delta_0$ that appears in Lemma 6.7 and that appears in our results is defined by

$$\delta_0 = \frac{1}{2} - \sigma_0, \tag{6.16}$$

so the Kim-Sarnak bound reads as $\delta_0 \geq \frac{25}{64}$.

Lemma 6.7 is stated for one place $v \in S_f$ but in Theorem 6.6 we wish to take advantage of the various places h is supported on. In order to do this, we will need to use Lemma 6.7 iteratively and the following abstract lemma in Hilbert space theory allows us to do so.

**Lemma 6.8.** *Let $G = G_1 \times G_2$ be a group acting unitarily on a Hilbert space $\mathcal{H}$. Let $K_i < G_i$ be subgroups, $g_i \in G_i$ be two given elements, and $F(g_i)$ two positive numbers satisfying the following statement: For each $i$, if $v, w \in \mathcal{H}$ are $K_i$-fixed vectors, then*

$$\langle g_i v, w\rangle \leq \|v\| \, \|w\| \, F(g_i).$$

*Then for any $v, w \in \mathcal{H}$ which are $K_1 \times K_2$-fixed we have that*

$$\langle g_1 g_2 v, w\rangle \leq \|v\| \, \|w\| \, F(g_1)F(g_2).$$

*Proof.* Let us denote for $i = 1, 2$ $V_i = \{v \in \mathcal{H} : v \text{ is } K_i\text{-fixed}\}$ and $U = V_1 \cap V_2$. Let $V_i'$ denote the orthocomplement of $U$ in $V_i$ and denote for a subspace $W$ of $\mathcal{H}$ by $P_W$ the orthogonal projection on $W$. We first note that $V_1, V_2$ are $K_2, K_1$-invariant respectively (because $K_1, K_2$ commute) and so the projections $P_{V_1}, P_{V_2}$ commute with the actions of $K_2, K_1$ respectively. It follows from here that given $v_1 \in V_1$, say, the projection $P_{V_2}(v_1)$ is fixed by both $K_1$ and $K_2$ i.e. $P_{V_2}(v_1) \in U$. This proves that $V_1'$ is orthogonal to $V_2$ or in a more symmetric manner, $V_1'$ is orthogonal to $V_2'$.

Let now $v, w$ be two $K_1 \times K_2$-fixed vectors. As $g_1 v$ is $K_2$-fixed, i.e. $g_1 v \in V_2$, we may write $g_1 v = P_U(g_1 v) + P_{V_2'}(g_1 v)$ and similarly $g_2 w = P_U(g_2 w) + P_{V_1'}(g_2 w)$. It follows that

$$\begin{aligned}\langle g_1 v, g_2 w\rangle &= \langle P_U(g_1 v) + P_{V_2'}(g_1 v), P_U(g_2 w) + P_{V_1'}(g_2 w)\rangle \\ &= \langle P_U(g_1 v), P_U(g_2 w)\rangle \leq \|P_U(g_1 v)\| \, \|P_U(g_2 w)\| \,. \end{aligned} \tag{6.17}$$

Let $\tilde{v} = \frac{P_U(g_1 v)}{\|P_U(g_1 v)\|}$. Then $\tilde{v}$ is $K_2$-fixed and so by the assumption of the lemma we conclude that

$$\|P_U(g_1 v)\| = \langle g_1 v, \tilde{v}\rangle \leq \|v\| \, F(g_1).$$

Similarly, $\|P_U(g_2 w)\| \leq \|w\| \, F(g_2)$. Plugging this into (6.17) yields

$$\langle g_1 v, g_2 w\rangle \leq \|v\| \, \|w\| \, F_1(g_1) F_2(g_2),$$

which is equivalent to the desired statement up to replacing $g_2$ by its inverse (note that the assumption on $g_i$ implies the corresponding assumption on $g_i^{-1}$). □

The final ingredient needed for the proof of Theorem 6.6 is the following

**Lemma 6.9.** *For each place $v \in S$ the group generated by $G_v$ and $K_{S_f}$ acts ergodically on $X_S$, that is, $\{w \in \mathcal{H} : w$ is both $G_v, K_{S_f}$-fixed$\}$ is the one dimensional space of constant functions.*

*Proof.* Let $Y_S = \mathrm{SL}_2(\mathbb{Z}[p^{-1} : p \in S_f]) \backslash \prod_{v \in S} \mathrm{SL}(\mathbb{Q}_v)$. The strong approximation property for $\mathrm{SL}_2$ implies that for any $v \in S$ the lattice $\mathrm{SL}_2(\mathbb{Z}[p^{-1} : p \in S_f])$ embeds densely in $\prod_{v' \in S \setminus \{v\}} \mathrm{SL}_2(\mathbb{Q}_{v'})$. This is equivalent to saying that $\mathrm{SL}_2(\mathbb{Q}_v)$ acts minimally on $Y_S$ (i.e. that any orbit is dense). In turn, this implies that $\mathrm{SL}_2(\mathbb{Q}_v)$ acts ergodically on $Y_S$ (by the duality trick for example). Now, consider the natural map $\psi : \mathrm{SL}_2 \to \mathrm{PGL}_2$. This map induces a map from $Y_S$ to $X_S$ (which we also denote by $\psi$) which intertwines the actions of $\mathrm{SL}_2(\mathbb{Q}_v)$ and $\psi(\mathrm{SL}_2(\mathbb{Q}_v)) < G_v$ on these spaces respectively. It follows that the action of $\psi(\mathrm{SL}_2(\mathbb{Q}_v))$ on $\psi(Y_S)$ is ergodic.

Let $w \in \mathcal{H}$ be a function on $X_S$ which is both $G_v$ and $K_{S_f}$-invariant. Its restriction to $\psi(Y_S)$ is constant by the ergodicity proved above. It follows that in order to show that $w$ is constant it is enough to show that the translates of $\psi(Y_S)$ by $K_{S_f}$ cover $X_S$. We briefly sketch the argument: There is a natural 'determinant map' $\det : G_S \to \prod_{v \in S} \mathbb{Q}_v^*/(\mathbb{Q}_v^*)^2$. Let us denote $\Delta = \prod_{v \in S} \mathbb{Q}_v^*/(\mathbb{Q}_v^*)^2$ and $\Delta' = \det(\Gamma_S) < \Delta$. It follows that there is a well defined map $\widetilde{\det} : \Gamma_S \backslash G_S = X_S \to \Delta' \backslash \Delta$. We leave it to the reader to show that the space $\psi(Y_S)$ is characterized as the preimage of the identity coset $\Delta'$ under $\widetilde{\det}$. Since det takes $K_{S_f}$ onto $\Delta' \backslash \Delta$, we conclude that indeed, translates of $\psi(Y_S)$ under $K_{S_f}$ cover $X_S$ as desired. □

*Proof of Theorem 6.6.* Let $\mathcal{H} = L^2(X_S, m_S)$ and for $v \in S$ let $\mathcal{H}_v$ be the orthocomplement to the $G_v$-fixed vectors. Let $\mathcal{H}_0 = \cap_{v \in S_f} \mathcal{H}_v$ and let $w_1, w_2 \in \mathcal{H}$ be as in the statement of the theorem. Write

$$w_i = P_{\mathcal{H}_0}(w_i) + P_{\mathcal{H}_0^\perp}(w_i),$$

and note that the decomposition $\mathcal{H} = \mathcal{H}_0 + \mathcal{H}_0^\perp$ is $G_S$-invariant. It follows that

$$\begin{aligned}\langle a_f(\mathrm{h})w_1, w_2\rangle &= \langle a_f(\mathrm{h})\left(P_{\mathcal{H}_0}(w_1) + P_{\mathcal{H}_0^\perp}(w_1)\right), P_{\mathcal{H}_0}(w_2) + P_{\mathcal{H}_0^\perp}(w_2)\rangle \\ &= \underbrace{\langle a_f(\mathrm{h})P_{\mathcal{H}_0}(w_1), P_{\mathcal{H}_0}(w_2)\rangle}_{(*)} + \underbrace{\langle a_f(\mathrm{h})P_{\mathcal{H}_0^\perp}(w_1), P_{\mathcal{H}_0^\perp}(w_2)\rangle}_{(**)}. \end{aligned} \tag{6.18}$$

Let us first argue that $(**) = \langle w_1, 1\rangle\langle 1, w_2\rangle$. The space $\mathcal{H}_0^\perp$ is the space generated by $\left\{\mathcal{H}_v^\perp\right\}_{v\in S_f}$. This implies that the vector $P_{\mathcal{H}_0^\perp}(w_1)$ is in the span of the vectors $P_{\mathcal{H}_v^\perp}(w_1)$ as $v$ runs through $S_f$. For each $v \in S_f$ the vector $P_{\mathcal{H}_v^\perp}(w_1)$ is both $G_v$ and $K_{S_f}$-fixed and so by Lemma 6.9 this implies that $P_{\mathcal{H}_v^\perp}(w_1) \in \mathcal{H}_c$, where $\mathcal{H}_c$ denotes here the 1-dimensional space of constant functions. We conclude that $P_{\mathcal{H}_0^\perp}(w_1) \in \mathcal{H}_c$, or in other words, $P_{\mathcal{H}_0^\perp}(w_1) = P_{\mathcal{H}_c}(w_1) = \langle w_1, 1\rangle$. Using this we see that

$$(**) = \langle\langle w_1, 1\rangle, P_{\mathcal{H}_0^\perp}(w_2)\rangle = \langle w_1, 1\rangle\langle 1, P_{\mathcal{H}_0^\perp}(w_2)\rangle. \tag{6.19}$$

In turn, $\langle P_{\mathcal{H}_0^\perp}(w_2), 1\rangle$ is the orthogonal projection of $P_{\mathcal{H}_0^\perp}(w_2)$ on $\mathcal{H}_c$, but as $\mathcal{H}_c \subset \mathcal{H}_0^\perp$, this projection equals $\langle w_2, 1\rangle$. We conclude from (6.19) that $(**) = \langle w_1, 1\rangle\langle 1, w_2\rangle$ as claimed.

We now analyze $(*)$ in (6.18). Because the decomposition $\mathcal{H} = \mathcal{H}_0 + \mathcal{H}_0^\perp$ is $G_S$-invariant the vectors $P_{\mathcal{H}_0}(w_1), P_{\mathcal{H}_0}(w_2)$ are fixed under $K_{S_f}, K^*$ respectively (where $K^*$ is as in the statement of the theorem). Order the primes in $S_f$ in some way $p_1 \dots p_k$ and denote $d_{p_i} = [K_{p_i} : K_{p_i}^*]$, so $d = [K_{S_f} : K^*] = \prod_{i=1}^k d_{p_i}$. We leave it to the reader to prove by a simple induction, using Lemmas 6.7, 6.8 that for $j = 1, \dots, k$

$$\langle\prod_{i=1}^j a_{p_i}(\mathrm{h})P_{\mathcal{H}_0}(w_1), P_{\mathcal{H}_0}(w_2)\rangle \ll_\epsilon \|P_{\mathcal{H}_0}(w_1)\|\,\|P_{\mathcal{H}_0}(w_1)\| \prod_{i=1}^j d_{p_j}^{\frac{1}{2}} \prod_{i=1}^k |\mathrm{h}|_{p_i}^{-\delta_0+\epsilon}. \tag{6.20}$$

In particular, for $j = k$ we obtain

$$(*) = \langle a_f(\mathrm{h})P_{\mathcal{H}_0}(w_1), P_{\mathcal{H}_0}(w_2)\rangle \ll_\epsilon \|w_1\|\,\|w_2\|\, d^{\frac{1}{2}}\, \mathrm{h}^{-\delta_0+\epsilon}. \tag{6.21}$$

Equations (6.21),(6.18) and the analysis carried above for $(**)$ now imply the validity of the theorem. □

6.4. **Tubes.** As explained in §6.2, we start with a $K_{S_f}$-invariant 'test function' $\varphi$ and we need to thicken the orbit $y_\omega L_\omega$ to a tube $\mathcal{T}$ and then apply (6.14) to the vectors $w_1 = \varphi, w_2 = \chi_{\mathcal{T}}$. In order for the use of (6.14) to be meaningful we need to control $d$ which is the index of the stabilizer of the tube in $K_{S_f}$. Also, the 'width' of the tube (i.e. the size of the thickening of the orbit) should be very small (at least in the real component) in order for the heuristics of §6.2 to take effect. This will hopefully motivate the constructions in this subsection.

**Definition 6.10.** Let $yL \subset X_S$ be a compact orbit of a closed subgroup $L < G_S$. Let $V = \oplus_{v\in S}V_v$ be a linear complement to $\mathrm{Lie}(L)$ in $\mathfrak{g}_S$. Let $U \subset V$ be a small enough open neighborhood of 0 so that the map $yL \times U \to X_S$ defined by $(z, u) \mapsto z\exp_S(u)$ is a homeomorphism onto its image and its image is open in $X_S$. The set $\mathcal{T}_U(yL) =$

$\{z \exp_S(u) : z \in yL, u \in U\}$ is called a *tube around the orbit $yL$ of width $U$*. We often denote the tube simply by $\mathcal{T}$. The width $U$ and the tube $\mathcal{T}$ are said to come from $V$.

A tube $\mathcal{T}_U(yL)$ gives us a coordinate system; a point of $\mathcal{T}$ can be written uniquely as $z \exp u$. We refer to $z$ as the *orbit coordinate* and to $u$ as the *width coordinate*. We shall need a few lemmas about tubes which we now turn to describe.

6.4.1. *Measures on tubes.* Given a tube $\mathcal{T} = \mathcal{T}_U(yL)$ around the compact orbit $yL$ coming from $V$, one could construct the following two natural probability measures supported on $\mathcal{T}$. The first is the normalized Haar measure $\frac{1}{m_S(\mathcal{T})} m_S|_{\mathcal{T}}$ which we will denote by $m_{\mathcal{T}}$. The second is the (pushforward of) the product measure $\eta \times m_U$ on $yL \times U \simeq \mathcal{T}$, where $\eta$ is the unique $L$-invariant probability measure on the orbit $yL$ and $m_U$ is the normalized restriction of the Haar measure on $V$ to $U$ (that is $m_U = \frac{1}{m_V(U)} m_V|_U$). We shall need to understand to some extent the connection between these two measures.

**Lemma 6.11.** *The measure $m_{\mathcal{T}}$ is absolutely continuous with respect to $\eta \times m_U$. Moreover, if we denote by $F(z,u)$ the Radon-Nikodym derivative; that is $dm_{\mathcal{T}} = F(z,u)d\eta(z)dm_U(u)$, then for $\eta$-almost any $z \in yL$, $\int_U F(z,u)dm_U(u) = 1$.*

*Proof.* The absolute continuity is left to be verified by the reader. As for the claim about the density $F$, we argue as follows. Let $\varphi(z) = \int_U F(z,u)dm_U(u)$. We will show that $\varphi$ is constant $\eta$-almost surely. As $\int_{yL} \varphi(z)d\eta(z) = m_{\mathcal{T}}(\mathcal{T}) = 1$ this constant must be equal to one.

Choose a fundamental domain $\mathcal{E}$ in $L$ for the orbit $yL$ and identify it with the orbit. Note that with this identification $\eta$ is just the restriction to $\mathcal{E}$ of a Haar measure[8] on $L$ scaled so that $\eta(\mathcal{E}) = 1$. Assume to get a contradiction that $\varphi$ is not constant $\eta$-almost surely. It follows that there are constants $c_2 < c_1$ so that the sets $E_1 = \{h \in \mathcal{E} : \varphi(yh) > c_1\}, E_2 = \{h \in \mathcal{E} : \varphi(yh) < c_2\}$ are of positive $\eta$-measure. There exists $h_0 \in L$ so that $\eta(E_1 \cap h_0^{-1}E_2) > 0$ and so if we let $\tilde{E}_1 = E_1 \cap h_0^{-1}E_2$ and $\tilde{E}_2 = h_0\tilde{E}_1$, then $\tilde{E}_i \subset \mathcal{E}$ are both of (the same) positive $\eta$-measure and differ from one another by left translation by $h_0$. The following calculation derives the desired contradiction:

$$\begin{aligned} c_1\eta(\tilde{E}_1) &\le \int_{\tilde{E}_1} \varphi(z)d\eta(z) \\ &= m_S(\tilde{E}_1 \exp_S(U)) = m_S(h_0\tilde{E}_1 \exp_S(U)) = m_S(\tilde{E}_2 \exp_S(U)) \qquad (6.22) \\ &= \int_{\tilde{E}_2} \varphi(z)d\eta(z) \le c_2\eta(\tilde{E}_2) = c_2\eta(\tilde{E}_1). \end{aligned}$$

□

Our aim now is to define the relevant family of tubes around the orbit $y_\omega L_\omega$ that will be of use to us. The first stage is to choose the correct linear complement from which the tubes will come.

[8] Note that $L$ must be unimodular, hence this measure is both left and right invariant.

6.4.2. *Choosing the linear complement.* When we come to argue the validity of Theorem 4.9(1),(2) for a given admissible radius h, we may replace the set of primes $S_f$ by the smallest set of primes on which h is supported on. Hence, without loss of generality we may (and will) assume that h is divisible by all the primes in $S_f$. We refer to such a radius h as having *full support.* The assumption that an admissible radius has full support is equivalent to the fact that the weak stable algebra of $a_f(\mathrm{h})$ attains the form

$$(\mathfrak{g}_S)^{\mathrm{ws}}_{a_f(\mathrm{h})} = \mathfrak{g}_\infty \oplus_{p\in S_f} \left\{ \begin{pmatrix} * & * \\ 0 & * \end{pmatrix} \in \mathfrak{g}_p \right\}. \tag{6.23}$$

**Definition 6.12.** Let $V = \oplus_{v\in S} V_v$ be defined as follows

$$V_\infty = \left\{ \begin{pmatrix} 0 & * \\ * & 0 \end{pmatrix} \in \mathfrak{g}_\infty \right\}; \quad \textit{For } p \in S_f,\ V_p = \left\{ \begin{pmatrix} * & * \\ 0 & * \end{pmatrix} \in \mathfrak{g}_p \right\}.$$

**Lemma 6.13.** *If the generalized branch $\mathcal{L}_\omega$ is nondegenerate then the subspace $V \subset \mathfrak{g}_S$ from Definition 6.12 is indeed a linear complement of* $\mathrm{Lie}(L_\omega)$ *which is contained in* $(\mathfrak{g}_S)^{\mathrm{ws}}_{a_f(\mathrm{h})}$ *for any admissible radius* h *of full support.*

*Proof.* The fact that $V \subset (\mathfrak{g}_S)^{\mathrm{ws}}_{a_f(\mathrm{h})}$ follows from the discussion preceding Definition 6.12. Recall that $L_\omega = A_\infty \times H_\omega$ where $H_\omega = \omega^{-1}\overline{\langle\gamma_x\rangle}_{G_{S_f}}\omega$ (see (6.5),(6.6)). Writing $\mathrm{Lie}(L_\omega) = \oplus_S \mathfrak{l}_v$ we see that $V_\infty$ indeed complements $\mathfrak{l}_\infty$. Let $\mathbb{T}$ be the algebraic subgroup of $\mathbb{G}$ defined as the Zariski closure of the group generated by $\gamma_x$. It is a one dimensional torus and $H_\omega$ is a compact open subgroup of the conjugation $\omega^{-1}\mathbb{T}(\prod_{S_f}\mathbb{Z}_p)\omega$. It follows that for any $v \in S_f$ the dimension of $\mathfrak{l}_v$ is 1 and so in order to argue that it complements $V_v$ we only need to argue that the inclusion $\mathfrak{l}_v \subset V_v$ does not hold. Such an inclusion would imply that there is a neighborhood of the identity in $H_\omega$ that consists of upper triangular matrices, which in turn would imply that a certain power of $\omega^{-1}\gamma_x\omega$ is upper triangular, contradicting the assumption that the generalized branch is nondegenerate. □

Henceforth, when speaking about a linear complement $V$ to $\mathrm{Lie}(L_\omega)$, we shall refer only to the subspace from Definition 6.12.

**Remark 6.14.** Because of the inclusion $V \subset (\mathfrak{g}_S)^{\mathrm{ws}}_{a_f(\mathrm{h})}$ (for any admissible radius h of full support), we conclude that if $U_0 \subset V$ is a small enough ball around zero, the conjugation $U_0^{a_f(\mathrm{h})}$ will be included in the domain of $\exp_S$. This implies that for any $U \subset U_0$ the identity $(\exp_S U)^{a_f(\mathrm{h})} = \exp_S\left(U^{a_f(\mathrm{h})}\right)$ holds. It follows that if $\mathcal{T}$ is a tube of width $U$ coming from $V$ around $y_\omega L_\omega$, then if the width $U$ is chosen within $U_0$, the pushed tube $\mathcal{T}a_f(\mathrm{h})$ satisfies

$$\mathcal{T}a_f(\mathrm{h}) = y_\omega L_\omega \exp_S(U) a_f(\mathrm{h}) = y_{\omega,\mathrm{h}} L_{\omega,\mathrm{h}} \exp_S\left(U^{a_f(\mathrm{h})}\right). \tag{6.24}$$

That is, $\mathcal{T}a_f(\mathrm{h})$ is a tube of width $U^{a_f(\mathrm{h})}$ around the compact orbit $y_{\omega,\mathrm{h}}L_{\omega,\mathrm{h}}$. Below, we will make the implicit assumption that all the widths considered are contained in the ball $U_0$.

6.4.3. *The tubes $\mathcal{T}_\omega^\delta$.* As explained above, we will need to construct tubes with shrinking real width component and with control on the subgroup of $K_{S_f}$ that stabilizes them. After describing this family of tubes we state a few lemmas that describe their relevant properties. The proofs of these lemmas will be postponed till after concluding the proof of Theorem 4.9.

Let us denote by $\tilde{B}$ a compact open subgroup of the group of upperr triangular elements in $K_{S_f}$ that lies in the domain of $\log_{S_f}$ and for which Remark 6.14 applies (that is, all conjugations $\tilde{B}^{a_f(\mathrm{h})}$ are in the domain of $\log_{S_f}$ for admissible radii h of full support). For $\delta > 0$ let $B_\delta^{V_\infty}$ be the ball of radius $\delta$ around 0 in the $\infty$-component of the linear complement $V$ from Definition 6.12.

**Lemma 6.15.** *There exists $\hat{\delta} > 0$ and an open compact subgroup $B = \prod_{S_f} B_p$ of $\tilde{B}$, such that for all $\delta < \hat{\delta}$, if we let $U^\delta = B_\delta^{V_\infty} \times \log_{S_f}(B)$, then for any $\omega \in K_{S_f}$ such that the generalized branch $\mathcal{L}_\omega$ is nondegenerate, the set $\mathcal{T}_\omega^\delta = y_\omega L_\omega \exp_S(U^\delta)$ is a tube around $y_\omega L_\omega$; that is, the map $y_\omega L_\omega \times U^\delta \to \mathcal{T}_\omega^\delta$ is a homeomorphism and the set $\mathcal{T}_\omega^\delta \subset X_S$ is open. Furthermore, the choice of $B, \hat{\delta}$ depends only on the original class $x$ and the set of places $S$ at hand. In particular, they are independent of $\omega$.*

**Lemma 6.16.** *Let $B, \hat{\delta}$ be as in Lemma 6.15 and let $\omega \in K_{S_f}$ be such that the generalized branch $\mathcal{L}_\omega$ is nondegenerate.*

*(1) There exists an open compact product subgroup $K^* = \prod_{S_f} K_v^* < K_{S_f}$ which stabilizes the tube $\mathcal{T}_\omega^\delta$ for any $\delta < \hat{\delta}$; that is $\mathcal{T}_\omega^\delta k = \mathcal{T}_\omega^\delta$ for any $k \in K^*, \delta < \hat{\delta}$. Moreover, if $x$ is non-split, we may choose $K^*$ to be independent of $\omega$.*

*(2) The measures $m_S(\mathcal{T}_\omega^\delta)$ satisfy $m_S(\mathcal{T}_\omega^\delta) \gg_{x,S_f,\mathcal{L}_\omega} \delta^2$. If $x$ is non-split, the implicit constant may be chosen to be independent of the generalized branch.*

## 6.5. **Concluding the main part of the proof.**

*Proof of parts*(1),(2) *of Theorem 4.9.* We follow the strategy presented in §6.2 and use freely all the notation introduced so far. Let h be an admissible radius and assume without loss of generality that it is of full support. Let $\varphi_0 \in L^2(X_\infty, m_\infty)$ be a $\kappa$-Lipschitz function. We let $\varphi = \varphi_0 \circ \pi$ be the lift of $\varphi_0$ to $X_S$. As $\int_{X_\infty} \varphi_0 dm_\infty = \int_{X_S} \varphi dm_S$, we see by Lemma 6.4 that part (1) of the theorem will follow once we prove

$$\left| \int_{X_S} \varphi d\eta_{\omega,\mathrm{h}} - \int_{X_S} \varphi dm_S \right| \ll_{x,S_f,\mathcal{L}_\omega,\epsilon} \max\left\{ \|\varphi\|_2, \kappa \right\} \mathrm{h}^{-\frac{\delta_0}{2}+\epsilon}. \tag{6.25}$$

Part (2) will follow once we establish that in the non-split case, the implicit constant in (6.25) does not depend on the generalized branch. Let $V < \mathfrak{g}_S$ be the linear complement from Definition 6.12. We apply Lemma 6.16 and use the notation introduced there to obtain a family of tubes $\mathcal{T}_\omega^\delta$ around $y_\omega L_\omega$ coming from $V$.

We denote $\mathcal{T}_{\omega,\mathrm{h}}^\delta$ the (pushed) tube $\mathcal{T}_\omega^\delta a_f(\mathrm{h})$ around the orbit $y_{\omega,\mathrm{h}} L_{\omega,\mathrm{h}}$, and $m_{\mathcal{T}_{\omega,\mathrm{h}}^\delta}$ the normalized restriction of $m_S$ to $\mathcal{T}_{\omega,\mathrm{h}}^\delta$. The width of[9] $\mathcal{T}_{\omega,\mathrm{h}}^\delta$ is $U^{\delta,\mathrm{h}} = (U^\delta)^{a_f(\mathrm{h})}$, where $U^\delta$ is

[9] The reader should not confuse the superscript $\delta$ with our notation for conjugation.

as in Lemma 6.15 (see Remark 6.14). We have

$$\left|\int_{X_S}\varphi d\eta_{\omega,\mathrm{h}}-\int_{X_S}\varphi dm_S\right| \leq \tag{6.26}$$
$$\underbrace{\left|\int_{X_S}\varphi d\eta_{\omega,\mathrm{h}}-\int_{X_S}\varphi dm_{\mathcal{T}^\delta_{\omega,\mathrm{h}}}\right|}_{(*)}+\underbrace{\left|\int_{X_S}\varphi dm_{\mathcal{T}^\delta_{\omega,\mathrm{h}}}-\int_{X_S}\varphi dm_S\right|}_{(**)}.$$

To estimate $(*)$ we define $\tilde{\varphi}_{\delta,\mathrm{h}}:\mathcal{T}^\delta_{\omega,\mathrm{h}}\to\mathbb{C}$ by $\tilde{\varphi}_{\delta,\mathrm{h}}(z\exp u)=\varphi(z)$ for $z\in y_{\omega,\mathrm{h}}L_{\omega,\mathrm{h}}, u\in U^{\delta,\mathrm{h}}$ and extend it to be zero outside the tube $\mathcal{T}^\delta_{\omega,\mathrm{h}}$ to obtain a function on $X_S$. By Lemma 6.11 it follows that

$$\int_{X_S}\tilde{\varphi}_{\delta,\mathrm{h}}dm_{\mathcal{T}^\delta_{\omega,\mathrm{h}}}=\int_{y_{\omega,\mathrm{h}}L_{\omega,\mathrm{h}}}\int_{U^{\delta,\mathrm{h}}}\varphi(z)F(z,u)dm_{U^{\delta,\mathrm{h}}}(u)d\eta_{\omega,\mathrm{h}}(z)=\int_{X_S}\varphi d\eta_{\omega,\mathrm{h}} \tag{6.27}$$

We therefore have the following estimate for $(*)$

$$\begin{aligned}(*) &= \left|\int_{X_S}\tilde{\varphi}_{\delta,\mathrm{h}}dm_{\mathcal{T}^\delta_{\omega,\mathrm{h}}}-\int_{X_S}\varphi dm_{\mathcal{T}^\delta_{\omega,\mathrm{h}}}\right|\\ &\leq \max\left\{|\varphi(y\exp(w))-\varphi(y)|:y\in y_{\omega,\mathrm{h}}L_{\omega,\mathrm{h}},w\in U^{\delta,\mathrm{h}}\right\}. \end{aligned}\tag{6.28}$$

Note that if we write $w\in U^{\delta,\mathrm{h}}$ as $(w_\infty,w_f)$, then for any $y\in y_{\omega,\mathrm{h}}L_{\omega,\mathrm{h}}$, $\varphi(y\exp_S(w))=\varphi_0(\pi(y)\exp_\infty(w_\infty))$ by the $K_{S_f}$-invariance of $\varphi$. As the maps induced by the actions of elements of the form $\exp_\infty(w_\infty)$, $\|w_\infty\|<1$ are all Lipschitz with some uniform Lipschitz constant $c_1$, the distance between $\pi(y)$ and $\pi(y)\exp_\infty(w_\infty)$ is $\leq c_1\delta$ and so by the Lipschitz assumption of $\varphi_0$ we obtain

$$(*)\leq c_1\kappa\delta. \tag{6.29}$$

We now estimate $(**)$. Let $\mathcal{H}=L^2(X_S,m_S)$ and denote

$$w_1=\varphi,\ w_2=\frac{1}{m_S(\mathcal{T}^\delta_\omega)}\chi_{\mathcal{T}^\delta_\omega}.$$

In order to appeal to Theorem 6.6 we observe that $w_1$ is $K_{S_f}$-fixed and $w_2$ is $K^*$-fixed, where $K^*$ is as in Lemma 6.16. By Lemma 6.16 the index $d=[K_{S_f}:K^*]$ depends only on $x,S_f$, and $\mathcal{L}_\omega$ and in the non-split case could be bounded by a number independent of the generalized branch. As for the norms, $\|w_1\|=\|\varphi\|$, and for $w_2$ we have $\|w_2\|=m_S(\mathcal{T}^\delta_\omega)^{-\frac{1}{2}}$. By Lemma 6.16 we have that $m_S(\mathcal{T}^\delta_\omega)\gg_{x,S_f,\mathcal{L}_\omega}\delta^2$ and so $\|w_2\|\ll_{x,S_f,\mathcal{L}_\omega}\delta^{-1}$. Furthermore, in the non-split case, the implicit constant can be taken to be independent of the generalized branch.

It now follows from Theorem 6.6 that

$$\begin{aligned}(**) &= \left|\langle\varphi,a_f(\mathrm{h})^{-1}\left(\frac{1}{m_S(\mathcal{T}^\delta_\omega)}\chi_{\mathcal{T}^\delta_\omega}\right)\rangle-\int_{X_S}\varphi dm_S\right| \\ &= |\langle a_f(\mathrm{h})w_1,w_2\rangle-\langle w_1,1\rangle\langle 1,w_2\rangle| \\ &\ll_{x,S_f,\mathcal{L}_\omega,\epsilon}\|\varphi\|_2\,\delta^{-1}\,\mathrm{h}^{-\delta_0+\epsilon},\end{aligned}\tag{6.30}$$

and that in the non-split case the implicit constant can be taken independent of the generalized branch. Combining (6.26),(6.29),(6.30), and choosing $\delta = c\,\mathrm{h}^{\frac{1}{2}(-\delta_0+\epsilon)}$ (the meaning of $c$ will become clear in a moment) we obtain (6.25) as desired (with $\epsilon$ replaced by $\frac{\epsilon}{2}$). Here the constant $c$ is chosen to protect us from the possible finitely many h's for which the inequality $\mathrm{h}^{\frac{1}{2}(-\delta_0+\epsilon)} < \hat{\delta}$ does not hold ($\hat{\delta}$ as in Lemma 6.15). Note that indeed, the constant $c$ depends only on $\hat{\delta}, S_f$, and $\epsilon$. By Lemma 6.15 we see that it actually depends on $x, S_f$, and $\epsilon$. This concludes the proof of Theorem 4.9. □

6.6. **Proofs of Lemmas 6.15,6.16.** We shall need the following auxiliary lemma which we leave without proof

**Lemma 6.17.** *There exists a neighborhood of the identity $W \subset G_S$, depending only on the class $x$, such that for any $\omega \in K_{S_f}$ and any $g \in W$, if $y_\omega L_\omega g \cap y_\omega L_\omega \neq \emptyset$ then $g \in L_\omega$.*

*Proof of Lemma 6.15.* The first restriction we impose on $\hat{\delta}$ is that it will be small enough so that in the real component, the map $(s,u) \mapsto \exp_\infty(s)\cdot\exp_\infty(u)$ from $B_{\hat{\delta}}^{\mathrm{Lie}(A_\infty)} \times B_{\hat{\delta}}^{V_\infty} \to G_\infty$ is a homeomorphism onto its open image. Choose $B = \prod_{S_f} B_p$ to be any product compact open subgroup of $\tilde{B}$ and define $U^\delta$ as in the statement of the lemma. At this stage we observe that for any $\delta < \hat{\delta}$ the map $L_\omega \times U^\delta \to G_S$ given by $(g,u) \mapsto g\exp_S(u)$ has an open image. To see this, note that the image is a product of open sets in each component: In the real component the image equals $A_\infty \cdot \exp_\infty B_\delta^{V_\infty}$ which is open by the choice of $\hat{\delta}$, while for any finite place $p \in S_f$, the $p$ component of the image is $(H_\omega)_p \cdot B_p$ which is easily seen to be open in the following way: Because of the fact that $V_p = \mathrm{Lie}(B_p)$ is a linear complement to $\mathrm{Lie}((H_\omega)_p)$, the product $(H_\omega)_p \cdot B_p$ clearly contains an open neighborhood of the identity in $G_p$. It now follows from the fact that both $(H_\omega)_p, B_p$ are groups, that their product is actually an open set.

The above establishes in particular, that the set $\mathcal{T}_\omega^\delta = y_\omega L_\omega \exp_S(U^\delta) \subset X_S$ is open. It follows that in order to conclude that $\mathcal{T}_\omega^\delta$ is indeed a tube around $y_\omega L_\omega$, we only need to argue the injectivity of the map $(z,u) \mapsto z\exp_S(u)$ from $y_\omega L_\omega \times U^\delta$ to $X_S$. We denote this map by $\psi_\omega$.

The second condition which we impose on $\hat{\delta}$ and on the choice of $B$ is that the product $\left(\exp_\infty(B_{\hat{\delta}}^{V_\infty})\right)^2 \cdot B^2 \subset W$, where $W$ is as in Lemma 6.17. Assuming the injectivity of $\psi_\omega$ fails, we obtain elements $u_\infty^{(i)} \in B_{\hat{\delta}}^{V_\infty}, b_i \in B$, $i = 1,2$ and a non-trivial intersection of the form

$$y_\omega L_\omega \exp_\infty(u_\infty^{(1)})b_1 \cap y_\omega L_\omega \exp_\infty(u_\infty^{(2)})b_2.$$

This shows that $y_\omega L_\omega \cap y_\omega L_\omega \exp_\infty(u_\infty^{(2)})\exp_\infty(-u_\infty^{(1)})b_2 b_1^{-1} \neq \emptyset$. It now follows from our choice of $\hat{\delta}$ and $B$ (by Lemma 6.17) that $\exp_\infty(u_\infty^{(2)})\exp_\infty(-u_\infty^{(1)}) \in A_\infty$ and that $b_2 b_1^{-1} \in H_\omega$. As $B$ is a group which intersects $H_\omega$ trivially (this is our assumption that the generalized branch $\mathcal{L}_\omega$ is nondegenerte), we conclude that $b_1 = b_2$. Furthermore, from the fact that $\mathrm{Lie}\,(A_\infty) \oplus V_\infty = \mathfrak{g}_\infty$, it is straightforward to deduce that if $\hat{\delta}$ is chosen small enough, then the inclusion $\exp_\infty(u_\infty^{(2)})\exp_\infty(-u_\infty^{(1)}) \in A_\infty$ implies that $u_\infty^{(1)} = u_\infty^{(2)}$. This establishes the injectivity of $\psi_\omega$ as desired □

*Proof of Lemma 6.16.* We first argue the validity of part (1). As pointed out in the proof of Lemma 6.15 above, if $\omega$ is such that $\mathcal{L}_\omega$ is nondegenerate, then the set $H_\omega \cdot B \subset G_{S_f}$ is open (here $B$ is as in Lemma 6.15). Moreover, as $B$ is compact, there exist a neighborhood of the identity in $G_{S_f}$, and in particular, a compact open subgroup $K^* = \prod_{S_f} K_p^*$, with the property that for any $k \in K^*$ we have $Bk \subset H_\omega B$. We now claim that for any tube $\mathcal{T}_\omega^\delta$ as in Lemma 6.15 we have $\mathcal{T}_\omega^\delta k = \mathcal{T}_\omega^\delta$. To argue the inclusion $\subset$ we note the following

$$\mathcal{T}_\omega^\delta k = y_\omega L_\omega \exp_\infty(B_\delta^{V_\infty})Bk \subset y_\omega L_\omega \exp_\infty(B_\delta^{V_\infty})H_\omega B = y_\omega L_\omega \exp_\infty(B_\delta^{V_\infty})B = \mathcal{T}_\omega^\delta.$$

The opposite inclusion follows by switching $k$ with $k^{-1}$.

If $x$ is non-spilt, it is not hard to see that the intersection $\cap_{\omega \in K_{S_f}}(H_\omega \cdot B)$ contains an open neighborhood around $e_f$. It then readily follows that this intersection contains an open neighborhood of $B$. We conclude similarly to the argument presented above that the group $K^*$ may be chosen to work for all the $\omega$'s simultaneously.

We briefly argue part (2) of the lemma. For each relevant $\omega$, it is not hard to see that the volume of the tube $m_S(\mathcal{T}_\omega^\delta)$ satisfies $c_1 m_{V_\infty}(B_\delta^{V_\infty}) \leq m_S(\mathcal{T}_\omega^\delta) \leq c_2 m_{V_\infty}(B_\delta^{V_\infty})$, where the constants $c_1, c_2$ are determined by the volume of the orbit $y_\omega L_\omega$ and the position of the linear space $V$, from which the width is coming, with respect to $\mathrm{Lie}(L_\omega)$. As the 2-dimensional volume $m_{V_\infty}(B_\delta^{V_\infty})$ is proportional to $\delta^2$, the claim regarding a single $\omega$ follows. In the non-split case, as the Lie algebras $\mathrm{Lie}(L_\omega)$ are uniformly transverse to $V$, the constant $c_1$ above can be taken to be uniform for all $\omega$ which finishes the proof. □

## 7. Proof of Lemma 5.1

We begin with a short discussion that will be helpful later on. Let $S_f$ be a finite set of primes. For an element $\delta \in G_{S_f}$ we denote $\Sigma_\delta = \overline{\langle\delta\rangle}_{G_{S_f}}$ and we say that $\delta$ is of compact type if $\Sigma_\delta$ is a compact group.

**Definition 7.1.** Let $\delta \in G_{S_f}$ be an element of compact type. Let us denote for any admissible radius $\mathrm{h}$ by $k_{\mathrm{h}}(\delta)$ the minimal positive integer $k$ for which $\delta^k$ belongs to the compact open subgroup $a_f(\mathrm{h})K_{S_f}a_f(\mathrm{h}^{-1})$.

Equivalently, $k_{\mathrm{h}}(\delta)$ is the order of (the image of) $\delta$ in the group $\Sigma_\delta/(\Sigma_\delta \cap a_f(\mathrm{h})K_{S_f}a_f(\mathrm{h}^{-1}))$ (note that this group is finite because $a_f(\mathrm{h})K_{S_f}a_f(\mathrm{h}^{-1})$ is open in $G_{S_f}$). It turns out that the behavior of the function $\mathrm{h} \mapsto k_{\mathrm{h}}(\delta)$ is essential for the proof of Lemma 5.1. We first state and prove the following:

**Lemma 7.2.** *Let $\delta \in G_{S_f}$ be an element of compact type and let $\mathrm{h} \mapsto k_{\mathrm{h}}(\delta)$ be the function defined above. Let $e_{\mathrm{h}}(\delta)$ be the positive number defined by the equation $k_{\mathrm{h}}(\delta) = e_{\mathrm{h}}(\delta)\,\mathrm{h}$. Then the function $\mathrm{h} \mapsto e_{\mathrm{h}}(\delta)$ attains only finitely many values and furthermore, if $\mathrm{h}_n$ is a divisibility sequence (i.e. $\mathrm{h}_{n+1} \,|\, \mathrm{h}_n$), then the sequence $e_{\mathrm{h}_n}(\delta)$ stabilizes.*

*Proof.* **Step 0**. Let us denote for an admissible radius $\mathrm{h}$ by $n_p(\mathrm{h})$ the integers satisfying $\mathrm{h} = \prod_{p\in S_f} p^{n_p(\mathrm{h})}$. We have the equality

$$a_f(\mathrm{h})K_{S_f}a_f(\mathrm{h}^{-1}) = \prod_{p\in S_f} a_p(p^{n_p(\mathrm{h})})K_p a_p(p^{-n_p(\mathrm{h})}),$$

and so the integer $k_{\mathrm{h}}(\delta)$ is the least common multiple (lcm) of the integers which we temporarily denote $j_p$, where $j_p$ is defined to be the minimal integer $j$ so that the $j$'th power of the $p$-component $\delta_p^j$ belongs to $a_p(p^{n_p(\mathrm{h})})K_p a_p(p^{-n_p(\mathrm{h})})$. A moment of thought shows that this implies that the statement of the lemma for a general finite set of primes $S_f$ follows from the corresponding statement for a single prime. Hence, from now on till the end of the proof we will assume that $S_f$ consists of a single prime $p$ and so the admissible radii that will be considered are positive powers of $p$. Let us denote $K_{p,n} = a_p(p^n)K_p a_p(p^{-n})$.

**Step 1** We claim that for all large enough $n$, $\Sigma_\delta \cap K_{p,n} < \Sigma_\delta \cap K_{p,(n-1)}$. To see this, note that the fact that $\delta$ is of compact type implies that when we represent an element $h \in \Sigma_\delta$ by a matrix in such a way that one of its entries is in $\mathbb{Z}_p$, then all the other entries must have $p$-adic absolute value bounded by some $p^{r_0}$ for some fixed $r_0$ (that depends on $\delta$). On the other hand

$$K_{p,n} = \left\{ \begin{pmatrix} a & p^{-n}b \\ p^n c & d \end{pmatrix} : a,b,c,d \in \mathbb{Z}_p, ad-bc \in \mathbb{Z}_p^* \right\}. \tag{7.1}$$

It follows that for $n > r_0$, given $h \in \Sigma_\delta \cap K_{p,n}$, when we write $h$ in the form given in (7.1) we must have that $b = pb'$ with $b' \in \mathbb{Z}_p$ and so if we denote $c' = pc$ then

$$h = \begin{pmatrix} a & p^{-(n-1)}b' \\ p^{(n-1)}c' & d \end{pmatrix},$$

and furthermore, $ad - bc = ad - b'c'$. This implies that $h \in K_{p,(n-1)}$ as desired.

The above shows that for all large enough $n$ there is an onto homomorphism

$$\Sigma_\delta/(\Sigma_\delta \cap K_{p,n}) \to \Sigma_\delta/(\Sigma_\delta \cap K_{p,(n-1)}),$$

which implies that $k_{p^{(n-1)}}(\delta)$ divides $k_{p^n}(\delta)$ for all large enough $n$. This divisibility relation will be needed in the next step.

**Step 2**. Let $\widehat{K}$ be a compact open subgroup of $G_p$ and let $k_0$ be the minimal positive integer $k$ so that $\delta^k \in \widehat{K}$. The following divisibility relation is straightforward

$$k_{p^n}(\delta^{k_0})|k_{p^n}(\delta)|k_0 k_{p^n}(\delta^{k_0}). \tag{7.2}$$

It is straightforward to show that (7.2) together with step 1 imply that the validity of the lemma for $\delta$ follows from its validity for $\delta^{k_0}$. This allows us to assume without loss of generality that we start with $\delta \in \widehat{K}$, where $\widehat{K}$ is some open compact subgroup of $G_p$ chosen at our convenience. We choose $\widehat{K}$ to be equal to the subgroup of $K_p$ consisting of elements having representatives which are matrices congruent to the identity modulo $p^2$.

The congruence relation modulo $p^2$ will be used towards the end of the proof. Moreover, the fact that we assume that $\delta \in K_p$ implies that $k_{p^n}(\delta)$ is the minimal positive integer

$k$ so that $\delta^k$ belongs to $K_p \cap K_{p,n}$. Let us denote $B_n = K_p \cap K_{p,n}$. A direct calculation shows

$$B_n = \left\{ \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in K_p : \frac{c}{p^n} \in \mathbb{Z}_p \right\}. \tag{7.3}$$

Similarly to step 1, the fact that $B_{n+1} < B_n$ implies the divisibility relation $k_{p^n}(\delta) | k_{p^{n+1}}(\delta)$ for any $n$.

**Step 3**. Working under the above assumptions on $\delta$, let $n_0$ be the maximal integer $n$ so that $\delta \in B_n$.

**Claim**: For any $n_0 \le n$ we have $k_{p^n}(\delta) = p^{n-n_0}$ and moreover, $\delta^{p^{n-n_0}} \in B_n \setminus B_{n+1}$.

Clearly, the validity of the above claim finishes the proof of the lemma. We prove it by induction on $n$. For $n = n_0$ the validity of the claim follows from the choice of $n_0$. Let us assume it holds for $n$. We know from step 2 that $k_{p^n}(\delta) | k_{p^{n+1}}(\delta)$ and from our inductive hypothesis saying that $\delta^{k_{p^n}(\delta)} \notin B_{n+1}$ that this divisibility relation is strict. It follows that $k_{p^{n+1}}(\delta) = j_0 p^{n-n_0}$ where $j_0$ is the minimal positive integer $j$ so that $\delta^{jp^{n-n_0}} \in B_{n+1}$, or said differently, such that the bottom left coordinate of $\delta^{jp^{n-n_0}}$ is divisible by $p^{n+1}$ in $\mathbb{Z}_p$. We will be finished once we show two things:

1. First, that $j_0 = p$ and so $k_{p^{n+1}}(\delta) = p^{n+1-n_0}$,
2. and second, that the bottom left coordinate of $\delta^{p^{n+1-n_0}}$ is not divisible by $p^{n+2}$ and so $\delta^{k_{p^{n+1}}(\delta)} \in B_{n+1} \setminus B_{n+2}$ which completes the inductive step.

Consider the sequence $\delta^{jp^{n-n_0}}$, $j = 1, 2, \ldots$ and denote

$$\delta^{jp^{n-n_0}} = \begin{pmatrix} a_j & b_j \\ c_j & d_j \end{pmatrix},$$

and note the recursive relation

$$c_{j+1} = c_1 a_j + c_j d_1. \tag{7.4}$$

We expand $c_1$ to a power series in $\mathbb{Z}_p$ and use the inductive assumption that $\delta^{p^{n-n_0}} \in B_n \setminus B_{n+1}$ and write

$$c_1 = m_1 p^n + m_2 p^{n+1} + u p^{n+2}, \tag{7.5}$$

where $m_1 \in \{1, 2, \ldots, p-1\}$, $m_2 \in \{0, 1, \ldots, p-1\}$, $u \in \mathbb{Z}_p$. We claim that for any $1 \le j$ we have

$$c_j = j m_1 p^n + j m_2 p^{n+1} + u_j p^{n+2} \text{ where } u_j \in \mathbb{Z}_p. \tag{7.6}$$

The validity of (1),(2) follows at once from (7.6) and the fact that $m_1 \in \{1, 2, \ldots p-1\}$. We prove the validity of (7.6) by induction on $j$. For $j = 1$, this is exactly (7.5). Now assume it holds for $j$ and write (using the congruence assumption on $\delta$ established in step 2)

$$a_j = 1 + p^2 A,\ d_j = 1 + p^2 D,\ A, D \in \mathbb{Z}_p.$$

Plugging this and (7.5),(7.6) into the recursive relation (7.4) we see that indeed

$$c_{j+1} = (j+1) m_1 p^n + (j+1) m_2 p^{n+1} + p^{n+2}(\ldots)$$

as desired. This completes the proof. □

**Remark 7.3.** It will be useful later on to note the following: A careful look at the argument giving Lemma 7.2 shows that for a fixed $\delta \in G_{S_f}$ of compact type, we have that there exists a positive constant $c$ such that $c \leq e_{\mathrm{h}}(\delta)$ for any admissible radius h, where $c$ depends only on two things:

(1) The power $k_0$ which we need to raise $\delta$ to so that each component $(\delta)_p^{k_0}$ will be in $K_p$ and congruent to the identity mod $p^2$.
(2) The maximal admissible radius $\mathrm{h} = \prod_{p \in S_f} p^{n_p}$ for which for any $p \in S_f$, $(\delta)_p^{k_0} \in B_{n_p}$ (this h measures how close $\delta^{k_0}$ is to being upper triangular).

Before turning to the proof of Lemma 5.1 we make yet another remark which will be used in the course of its proof.

**Remark 7.4.** Given a class $x \in X_\infty$ with a periodic $A_\infty$-orbit and a representative $\Lambda_x \in x$, then the matrix $a_\infty(t_x) = \mathrm{diag}\left(e^{\frac{t_x}{2}}, e^{-\frac{t_x}{2}}\right)$ stabilizes the lattice $\Lambda_x$ (that is, as a subset of $\mathbb{R}^2$, $\Lambda_x = \Lambda_x a_\infty(t_x)$). As $\Lambda_x$ is a lattice, it follows that $a_\infty(t_x)$ is conjugate to an integer matrix and so its eigenvalues $e^{\pm\frac{t_x}{2}}$, are algebraic integers of degree 2. The quadratic extension $\mathbb{F}_x$ from Remark 4.5 is the one generated by them. As these eigenvalues are Galois conjugates whose product is equal to 1, we conclude furthermore that they are units in the ring of integers of $\mathbb{F}_x$ and as such, by Dirichlet's unit theorem, they are integer powers of the fundamental unit of this field. In fact, the term "fundamental unit" will sometimes be used here to refer to the unit in the ring of integers which is of absolute value $> 1$ and which generates the group of totally positive units (i.e. those units both of whose embeddings into the reals are positive). If the fundamental unit is $\epsilon = e^{\frac{t_0}{2}}$, then the reader will easily verify that the image $\Lambda_x a_\infty(t_0)$ is contained in the $\mathbb{Q}$-span of $\Lambda_x$. This shows that if we write $x = \Gamma_\infty g_x$, then there is a rational matrix $\delta_x$ which solves $\delta_x g_x = g_x a_\infty(t_0)$ and in fact, $t_x = k t_0$ where $k$ is the minimal positive integer such that $\delta_x^k$ is an integer matrix.

*Proof of Lemma 5.1.* We first argue part (1) of the lemma. A short counting argument shows that the cardinality of the sphere $\mathcal{S}_{\mathrm{h}}(x)$ is proportional to h (were the proportionality constant depends on $S_f$). For each $x'$ on the sphere, let $s_{x'}$ be the minimal positive number such that $x' a_\infty(s)$ returns to the sphere. The total length is then $\mathbf{t}_x(\mathrm{h}) = \sum_{x' \in \mathcal{S}_{\mathrm{h}}(x)} s_{x'}$. We will show below that for any $x' \in \mathcal{S}_{\mathrm{h}}(x)$ $s_{x'} \leq t_x$. This will establish the inequality $\mathbf{t}_{\mathrm{h}}(x) \ll_{x,S_f} \mathrm{h}$ which is half of of the statement in part (1) of the lemma. The other half, namely the inequality $\mathrm{h} \ll_{x,S_f} \mathbf{t}_x(\mathrm{h})$, actually follows from part (2) of the lemma.

Let $x' \in \mathcal{S}_{\mathrm{h}}(x)$ be given. By Lemma 4.1 we see that there exists $y \in \pi^{-1}(x)$ such that $x' = \pi(y a_f(\mathrm{h}))$. As $\pi$ intertwines the $A_\infty$-actions on $X_S, X_\infty$ we see that $x = x a_\infty(t_x) = \pi(y a_\infty(t_x))$ and so if we let $\tilde{y} = y a_\infty(t_x)$ then $\tilde{y} \in \pi^{-1}(x)$ and again by Lemma 4.1 we have that $x'' = \pi(\tilde{y} a_f(\mathrm{h})) \in \mathcal{S}_{\mathrm{h}}(x)$. The following calculation then shows that indeed $s_{x'} \leq t_x$ as was claimed:

$$x' a_\infty(t_x) = \pi(y a_f(\mathrm{h}) a_\infty(t_x)) = \pi(y a_\infty(t_x) a_f(\mathrm{h})) = \pi(\tilde{y} a_f(\mathrm{h})) = x'' \in \mathcal{S}_{\mathrm{h}}(x).$$

We now turn to argue part (2) of the lemma. Let $\mathcal{L} = \mathcal{L}_{g_x,\omega}$ be a nondegenerate generalized branch of $\mathcal{G}_S(x)$ (here $x = \Gamma_\infty g_x$ and $\omega \in K_{S_f}$). Let $\mathbb{F}_x$ be the quadratic field from which the periodic orbit $xA_\infty$ arises from and let $t_0 > 0$ be such that $e^{\frac{t_0}{2}}$ is the fundamental unit of $\mathbb{F}_x$ as in Remark 7.4. In the notation of the same remark, let $\delta_x$ be the rational matrix satisfying $\delta_x g_x = g_x a_\infty(t_0)$. We replace the set of places $S$ by a bigger set if necessary, $\widetilde{S}$, so that $\delta_x \in \Gamma_{\widetilde{S}}$. We then consider the bigger graph $\mathcal{G}_{\widetilde{S}}(x)$ which contains the original graph and we further consider its following generalized branch: Write $\widetilde{S}_f = S_f \cup T$ and define $\widetilde{\omega} \in K_{\widetilde{S}_f}$ to be identical to $\omega$ in the components corresponding to the primes in $S_f$ and equal the identity in the components corresponding to primes in $T$. We then define $\widetilde{\mathcal{L}}$ to be the generalized branch $\mathcal{L}_{g_x,\widetilde{\omega}}$ of $\mathcal{G}_{\widetilde{S}}(x)$. Note that because of the way we defined $\widetilde{\omega}$, the generalized branch $\widetilde{\mathcal{L}}$ is nondegenerate as well.

Denote as before by $x_{\mathrm{h}}$ the class in $\widetilde{\mathcal{L}} \cap \mathcal{S}_{\mathrm{h}}(x)$. We are interested in analyzing the length $t_{x_{\mathrm{h}}}$ of the orbit $x_{\mathrm{h}} A_\infty$. By Remark 7.4, there exists a positive integer $\widehat{k}_{\mathrm{h}}$ satisfying

$$t_{x_{\mathrm{h}}} = \widehat{k}_{\mathrm{h}} t_0. \tag{7.7}$$

In fact, for later purposes, note that in our discussion $x$ and the representative $g_x$ are fixed but we will play with the branch later on, i.e. with the choice of $\widetilde{\omega}$ (which in our setting is defined by $\omega$), and so we should actually record the dependency in $\widetilde{\omega}$ in our notation and denote $\widehat{k}_{\mathrm{h}}(\widetilde{\omega})$. The function $k_{\mathrm{h}}(\cdot)$ from definition 7.1 and $\widehat{k}_{\mathrm{h}}(\cdot)$ are closely related as will be seen below.

The number $\widehat{k}_{\mathrm{h}}(\widetilde{\omega})$ is by definition the minimal positive integer such that $x_{\mathrm{h}} a_\infty(kt_0) = x_{\mathrm{h}}$ or, if we prefer working in the extension $X_{\widetilde{S}}$, it is the minimal positive integer so that $\Gamma_{\widetilde{S}}(g_x, \widetilde{\omega}) a_f(\mathrm{h}) a_\infty(kt_0)$ returns to the fiber $\pi^{-1}(x_{\mathrm{h}})$. Because of the identity $\delta_x g_x = g_x a_\infty(t_0)$ and the fact that $\delta_x \in \Gamma_{\widetilde{S}}$ we see that $\Gamma_{\widetilde{S}}(g_x, \widetilde{\omega}) a_f(\mathrm{h}) a_\infty(kt_0) = \Gamma_{\widetilde{S}}(g_x, \delta_x^{-k} \widetilde{\omega} a_f(\mathrm{h}))$, and so this point lies in the same fiber as $\Gamma_{\widetilde{S}}(g_x, \widetilde{\omega} a_f(\mathrm{h}))$ (i.e. above $x_{\mathrm{h}}$) if and only if the quotient $a_f(\mathrm{h}^{-1}) \widetilde{\omega}^{-1} \delta_x^k \widetilde{\omega} a_f(\mathrm{h})$ belongs to $K_{\widetilde{S}_f}$ (see Remark 3.1). That is, $\widehat{k}_{\mathrm{h}}(\widetilde{\omega})$ is the minimal positive integer $k$ for which the $(\widetilde{\omega}^{-1} \delta_x \widetilde{\omega})^k \in a_f(\mathrm{h}) K_{\widetilde{S}_f} a_f(\mathrm{h}^{-1})$. This establishes the equality

$$k_{\mathrm{h}}(\delta_x^{\widetilde{\omega}}) = \widehat{k}_{\mathrm{h}}(\widetilde{\omega}).$$

The validity of part (2) of the lemma now follows immediately from Lemma 7.2 and (7.7) which together imply $c_{\mathcal{L}_{g_x,\omega}}(\mathrm{h}) = t_0 e_{\mathrm{h}}(\delta_x^{\widetilde{\omega}})$.

As for part (3) we argue as follows: Assume first that $x$ is non-split with respect to $S_f$. As noted in Remark 7.3 the lower bound for the function $\mathrm{h} \mapsto e_{\mathrm{h}}(\delta_x^{\widetilde{\omega}})$, which gives us the lower bounds for the functions $c_{\mathcal{L}_{g_x,\omega}}(\mathrm{h})$, depends only on two things:

(1) The smallest power $k_0$ for which $\delta_x$ belongs to the subgroup of $K_{\widetilde{S}_f}$ consisting of elements congruent to the identity modulo $p^2$ in each component (note that we may ignore the conjugation by $\widetilde{\omega}$ as this is a normal subgroup of $K_{\widetilde{S}_f}$).

(2) The $p$-adic norms $|c_p|_p$, where $c_p$ is the left bottom coordinate of the $p$-component of $(\widetilde{\omega}^{-1} \delta_x \widetilde{\omega})^{k_0}$ where $p \in \widetilde{S}_f$.

It is clear that $k_0$ depends only on $x$ and the original set of primes $S_f$ and does not vary with $\omega$ (i.e. with the generalized branch). Also, for primes $p \in S_f$, the $p$-adic norm $\left|c_p\right|_p$ is bounded from below as $\omega$ ranges over $K_{S_f}$ because $x$ is non split. Finally, for the primes $p \in \widetilde{S}_f \setminus S_f$, as the $p$'th component of $\widetilde{\omega}$ equals the identity, the $p$'th component of $(\widetilde{\omega}^{-1}\delta_x\widetilde{\omega})^{k_0}$ is independent of $\omega$. We conclude that

$$\inf\left\{c_{\mathcal{L}_{g_x,\omega}}(\mathrm{h}) : \omega \in K_{S_f}, \mathrm{h} \text{ is an admissible radius}\right\} > 0$$

as desired. We leave it as an exercise to the reader to show that in the split case this infimum equals zero.

□

## 8. Applications to continued fractions

In this section we will prove the results stated in §2.3. The main goal is the deduction of Theorem 2.1 from Theorem 2.8. Throughout the rest of this paper we use slightly different notation than that previously introduced. We let $G = \mathrm{PSL}_2(\mathbb{R})$, $\Gamma = \mathrm{PSL}_2(\mathbb{Z})$, and $X = \Gamma\backslash G$. We denote the projection from $G$ to $X$ by $\pi$. The group $G$ acts on the upper half plane $\mathbb{H} = \{z = x + iy : y > 0\}$ by Möbius transformations; if $g$ is represented by a matrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}, \tag{8.1}$$

then for $z \in \mathbb{H}$, $gz = \frac{az+b}{cz+d}$. This action preserves the hyperbolic metric $ds^2 = \frac{dx^2+dy^2}{y^2}$ and so induces an action of $G$ on the unit tangent bundle $T^1\mathbb{H}$. The action of $G$ on $T^1\mathbb{H}$ is free and transitive hence allows us to identify $G$ with $T^1\mathbb{H}$ once we choose a base point. We make the usual choice of the base point to be the tangent vector pointing upwards through $i \in \mathbb{H}$. With this identification the geodesic flow on $G = T^1\mathbb{H}$ corresponds to the action from the right of the positive diagonal subgroup

$$A = \{a(t)\} = \left\{\mathrm{diag}\left(e^{t/2}, e^{-t/2}\right) : t \in \mathbb{R}\right\} < G.$$

**Remark 8.1.** The reason we chose to work in previous sections with the group $\mathrm{PGL}_2$ rather than with $\mathrm{PSL}_2$ which fits better for our application is that if one works with $\mathrm{PSL}_2$, equation (4.11), which lies at the heart of our arguments, needs to be adjusted. The diagonal matrix $\mathrm{diag}\,(q, 1)$ is no longer an element of the group (and so not an element of the lattice) and needs to be replaced with $\mathrm{diag}\,(q, q^{-1})$. This will then produce a slightly weaker relation, namely it will relate the geodesic loop corresponding to a quadratic irrational $\alpha$ to the one corresponding to $q^2\alpha$ (rather than $q\alpha$). This would have allowed us to prove only equidistribution along sequences of the form $q_n^2\alpha$. Note however that what makes it possible for us to apply the results of previous sections to the above settings is that the natural map from $\mathrm{PSL}_2(\mathbb{Z})\backslash \mathrm{PSL}_2(\mathbb{R})$ to $\mathrm{PGL}_2(\mathbb{Z})\backslash \mathrm{PGL}_2(\mathbb{R})$ is one to one and onto. This is not the case when one replaces $\mathbb{R}, \mathbb{Z}$ with other rings.

In the following subsections we briefly recall the connection between the geodesic flow on $X$ and continued fractions. As mentioned in the introduction, this connection was discovered by Artin in [Art82] and described in great detail in [Ser85]. The reader who is unfamiliar with this material is advised to consult [EW11, §9.6] as we shall follow the exposition presented there.

8.1. **Cross-sections.** We now wish to introduce the notion of a cross-section. We are being rather restrictive below as we only want to discuss a specific example hence we see no use in greater generality. Given a Borel measurable set $C \subset X$, we let $r_C : C \to \mathbb{R}_{\geq 0} \cup \{\infty\}$ be defined by $r_C(x) = \inf \{t > 0 : xa(t) \in C\}$. The function $r_C$ is called *the return time* function to $C$. The set $C$ is called a *cross-section* for $a(t)$ if the return time functions for positive and negative times are bounded from below by some fixed positive number and the map $(x, t) \mapsto xa(t)$ from $\{(x, t) : x \in C, 0 \leq t < r_C(x)\} \to X$ is a measurable isomorphism onto its image in $X$. The *first return map* $T_C$ is defined to be $T_C(x) = xa(r_C(x))$, where this makes sense; i.e. for $x$ belonging to $\{x \in C : r_C(x) < \infty\}$. In fact, we will be interested only in points which return infinitely often in the future and past to $C$, thus we define the domain of the first return map to be

$$\begin{aligned}\mathrm{Dom}_{T_C} = \{x \in C : &\text{ there are infinitely many} \\ &\text{positive and negative } t\text{'s with } xa(t) \in C\}.\end{aligned} \tag{8.2}$$

Note that $T_C : \mathrm{Dom}_{T_C} \to \mathrm{Dom}_{T_C}$ is invertible.

We now wish to define the relevant cross-section for the geodesic flow in $X$. An element $g \in G$ represented by a matrix as in (8.1) corresponds to a tangent vector of unit length to the upper half plane. It then defines a geodesic in $\mathbb{H}$ which hits the boundary of $\mathbb{H}$ in two points. We denote the *endpoint* and *startpoint* of the geodesic it defines by $e_+(g), e_-(g)$ respectively. Clearly we have $e_+(g) = \frac{a}{c}, e_-(g) = \frac{b}{d}$, where we allow $\infty$ as a possible value. Any element $g \in G$ has a unique decomposition (the Iwasawa decomposition) of the form

$$g = n(t)a(s)k_\theta = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} e^{s/2} & 0 \\ 0 & e^{-s/2} \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \tag{8.3}$$

where $t, s \in \mathbb{R}$, and $\theta \in [0, \pi)$. The notation $n(t), a(s), k_\theta$ should be understood from (8.3). An element $g$ having the above decomposition corresponds to the tangent vector to the point $t + ie^s \in \mathbb{H}$ of angel $2\theta$ in the *clockwise* direction from the vector pointing upwards. Consider the following sets:

$$\begin{aligned}\mathcal{C}^+ &= \{g = a(s)k_\theta \in G : e_+(g) \in (0, 1), e_-(g) < -1\}; \\ \mathcal{C}^- &= \{g = a(s)k_\theta \in G : e_+(g) \in (-1, 0), e_-(g) > 1\}; \\ \mathcal{C} &= \mathcal{C}^+ \cup \mathcal{C}^-.\end{aligned} \tag{8.4}$$

The set $\mathcal{C}$ consists of those tangent vectors whose base-point lies on the imaginary axis with some restriction on the angle $\theta$ related to the height $e^s$ of the base point. It should be clear from the geometric picture described above that the range of 'allowed angles' for such a tangent vector, say in $\mathcal{C}^+$, is a subinterval of $(\frac{\pi}{4}, \frac{\pi}{2})$ with $\frac{\pi}{2}$ being its right-end-point. In §9 we will workout these intervals exactly. We denote the sets $\pi(\mathcal{C}), \pi(\mathcal{C}^+), \pi(\mathcal{C}^-)$ by $C, C^+, C^-$ respectively. The following lemma is proved in [EW11, §9.6].

**Lemma 8.2.** *The following hold*

*(1) The set $\mathcal{C}$ injects into $X$ under $\pi$; that is, for each $x \in C$ corresponds a unique $g \in \mathcal{C}$ with $\pi(g) = x$.*
*(2) The set $C$ is a cross-section for the geodesic flow on $X$.*
*(3) The domain of $T_C$ corresponds to those $g \in \mathcal{C}$ for which both $e_+(g), e_-(g)$ are irrational.*
*(4) For $g \in \mathcal{C}^+$, if $T_C(\pi(g))$ is defined, then $T_C(\pi(g)) \in C^-$. An analogue statement with $+$ replaced by $-$ holds.*

It will be convenient for us to introduce a 'thickening' of the cross-section $C$ which will denoted by $B$. The following lemma is left to be verified by the reader.

**Lemma 8.3.** *There exists a constant $\epsilon_0 > 0$ (which will be fixed throughout) such that the following statements hold*

*(1) The the map $(g, t) \mapsto ga(t)$ from $\mathcal{C} \times (0, \epsilon_0)$ to the set*
$$\mathcal{B} = \{ga(t) : g \in \mathcal{C}, t \in (0, \epsilon_0)\} \tag{8.5}$$
*is one to one and onto, and the set $\mathcal{B}$ is open in $G$.*
*(2) Let $B = \pi(\mathcal{B})$. The restriction $\pi : \mathcal{B} \to B$ is one to one and onto and the set $B \subset X$ is open.*

The constant $\epsilon_0$ introduced in the above lemma is a lower bound for the return time function, $r_C$, to the cross-section $C$. The importance of part (2) of the above lemma is that it gives us a well defined way of lifting points in $X$ near the cross-section to the group $G$ in which it is more convenient to work. The combination of parts (1) and (2) gives us natural coordinates on $B$; any point $x \in B$ can be written uniquely as $x_C a(t)$ where $x_C \in C$ and $t \in (0, \epsilon_0)$.

In our discussion we will encounter certain measures on the cross-section $C$ which are invariant under the first return map and we will need a procedure to construct from them measures on the ambient space $X$ which are invariant under the geodesic flow; that is, under the action of the group $A$.

Let $\tilde{\mu}$ be a probability measure on $C$. We define the *suspension* of $\tilde{\mu}$ to be the measure $\sigma_{\tilde{\mu}}$ on $X$ which is given by the following rule of integration: For $f \in C_c(X)$
$$\int_X f(x) d\sigma_{\tilde{\mu}}(x) = \int_C \int_0^{r_C(x)} f(xa(t)) dt d\tilde{\mu}(x). \tag{8.6}$$

**Lemma 8.4.** *If $\tilde{\mu}(\mathrm{Dom}_{T_C}) = 1$ and $\tilde{\mu}$ is $T_C$-invariant, then the suspension $\sigma_{\tilde{\mu}}$ is $A$-invariant. Furthermore, $\sigma_{\tilde{\mu}}(X) = \int_C r_C d\tilde{\mu}$.*

*Proof.* This is follows from [EW11, Lemma 9.23] taking into account that $T_C$ is invertible on $\mathrm{Dom}_{T_C}$. □

**Definition 8.5.** Given a function $f : C \to \mathbb{C}$, we denote by $\widehat{f} : X \to \mathbb{C}$ the following function

$$\widehat{f}(x) = \begin{cases} f(x_C) & \text{if } x \in B \text{ has coordinates } (x_C, t), \\ 0 & \text{if } x \notin B \end{cases}$$

Note that with the above definition, given a measure $\tilde{\mu}$ on $C$ and a function $f : C \to \mathbb{C}$, equation (8.6) translates to the following useful formula which will be used frequently below

$$\int_X \widehat{f} d\sigma_{\tilde{\mu}} = \epsilon_0 \int_C f d\tilde{\mu}. \tag{8.7}$$

8.2. **The Gauss map.** Let $I = (0,1)$ and $S : I \to I$ be the Gauss map; i.e. the map defined by the formula $S(y) = \frac{1}{y} - \lfloor \frac{1}{y} \rfloor$. Note that strictly speaking $S(y)$ is not in $I$ for points of the form $y = \frac{1}{m}$. The reader will easily verify that $S^n(y)$ is well defined for all positive $n$ if and only if $y$ is irrational. This slight inconvenience will not bother us as we will only apply the Gauss map to irrational points. Let $I_{\text{irr}} = I \setminus \mathbb{Q}$. Consider the following subsets of $\mathbb{R}^2$:

$$D = \left\{ (y,z) : y \in I, 0 < z < \frac{1}{1+y} \right\}, \ D_{\text{irr}} = \left\{ (y,z) \in D : y \in I_{\text{irr}} \right\}. \tag{8.8}$$

Let $\bar{S} : D \to D$ be the map given by $\bar{S}(y,z) = (S(y), y(1-yz))$ and note similarly that strictly speaking, in order to iterate $\bar{S}$ as many times as we wish we need to restrict to points in $D_{\text{irr}}$. Recall (see for example [EW11, §3.4]) that the normalized restriction of the Lebesgue measure on $\mathbb{R}^2$ to $D$, which we denote here by $\lambda$, is an $\bar{S}$-invariant probability measure. This is the so called *invertible*[10] *extension of the Gauss map* as when one projects on the first coordinates, one recovers the Gauss map and the Gauss-Kuzmin measure $\nu$ introduced in the introduction. That is if $p : D \to I$ denotes the projection on the first coordinate, then

$$p_* \lambda = \nu. \tag{8.9}$$

As we will see below, the dynamical system $\bar{S} : D \to D$ can basically be identified with $T_C : C \to C$.

8.3. **Relation to the Gauss map.** Consider the maps $\tau_+ : C^+ \to D, \tau_- : C^- \to D$ defined by the following formulas: For $x = \pi(g) \in C$, where $g \in \mathcal{C}$ is of the form (8.1):

$$\text{For } g \in \mathcal{C}^+, \ \tau_+(x) = (e_+(g), \frac{1}{e_+(g) - e_-(g)}) = (\frac{a}{c}, cd), \tag{8.10}$$
$$\text{For } g \in \mathcal{C}^-, \ \tau_-(x) = (-e_+(g), \frac{1}{-e_+(g) + e_-(g)}) = (-\frac{a}{c}, -cd).$$

We let $\tau : C \to D$ be the union of $\tau_+$ and $\tau_-$. The formulas in (8.10) can be stated geometrically as follows: For a tangent vector $g \in \mathcal{C}$ and $x = \pi(g)$, $\tau(x) = (y,z) \in D$, where $y$ is the absolute value of the end point of the semicircle corresponding to $g$ and

[10] The term 'invertible' refers to the fact that when restricted to a subset of $D$, $\bar{S}$ is indeed invertible. This subset is obtained by neglecting a certain set of Lebesgue measure zero (see [EW11, Prop. 3.15]).

$z^{-1}$ is the diameter of it. For any endpoint $y \in (0,1)$ (resp. $y \in (-1,0)$) and any diameter $z^{-1} > 1$, we can attach a well defined semicircle in $\mathbb{H}$ which corresponds to a unique point in $\mathcal{C}^+$ (resp. $\mathcal{C}^-$). This shows that $\tau$ is one to one and onto (and in fact, a homeomorphism) from $C^+$ (resp. $C^-$) to $D$ which is the area below the graph of the function $y \mapsto (1+y)^{-1}$. The following basic lemma is proved in [EW11, §9.6]. It establishes the link between the geodesic flow and the Gauss map.

**Lemma 8.6.** *The following diagram commutes (for points $x \in C$ for which $T_C(x)$ is defined)*

$$\begin{tikzcd} C \arrow[r, "T_C"] \arrow[d, "\tau"'] & C \arrow[d, "\tau"] \\ D \arrow[r, "\bar{S}"'] & D. \end{tikzcd}$$

Note that as $\tau : C \to D$ is 'almost' an isomorphism (it is two to one), and so the above lemma basically says that any dynamical question about the dynamical system $\bar{S} : D \to D$ can be pulled to a dynamical question on $T_C : C \to C$. In our case the dynamical question is that of equidistribution of certain $\bar{S}$-invariant measures. Using the suspension construction we will see that the equidistribution questions for the dynamical system $T_C : C \to C$ translate to equidistribution questions of certain $A$-invariant measures on $X$.

8.4. **Reducing the statement of Theorem 2.1 to the cross-section.** We will be interested in two types of measures on the cross-section $C$ defined above. The first is the following version of the Lebesgue measure: We use $\tau_+$ (resp. $\tau_-$) to pull the (normalized restriction of) Lebesgue measure $\lambda$ from $D$ to $C^+$ (resp. $C^-$) and denote the resulting measure by $\tilde{\lambda}^+$ (resp. $\tilde{\lambda}^-$). Further denote $\tilde{\lambda} = \frac{1}{2}\tilde{\lambda}^+ + \frac{1}{2}\tilde{\lambda}^-$. Clearly $\tilde{\lambda}$ is $T_C$-invariant and $\tau_*(\tilde{\lambda}) = \lambda$.

The second type of measures on $C$ are those coming from quadratic irrationals. We recall some notation introduced earlier. Let $\alpha$ be a quadratic irrational. Let $g_\alpha$ be as in (2.5). We chose to define $g_\alpha$ as we did so as to ensure that its determinant is positive and hence it corresponds naturally to an element of $G$ with endpoint $\alpha$. Let $x_\alpha \in X$ be the corresponding point (that is $x_\alpha = \pi(\frac{1}{\sqrt{\det g_\alpha}} g_\alpha)$) and $\mu_\alpha$ the $A$-invariant probability measure supported on the periodic orbit $x_\alpha A = \{x_\alpha a(t) : t \in [0, t_\alpha)\}$, where $t_\alpha$ is the length of the orbit (see Lemma 4.4). We claim that the intersection $C \cap x_\alpha A$ is a non-empty finite set contained in $\mathrm{Dom}_{T_C}$. In fact, any geodesic in the upper half plane that corresponds to a semi-circle, projects to a set in $X$ that intersects $C$ non-trivially. By Lemma 8.2(3), if the end points of the geodesic are irrational, the intersection is in $\mathrm{Dom}_{T_C}$. Finally, the finiteness follows from the fact that $C$ is a cross-section together with the fact that the orbit $x_\alpha A$ is of finite length.

Let us denote by $\tilde{\mu}_\alpha$ the normalized counting measure on $C \cap x_\alpha A$. Clearly $\tilde{\mu}_\alpha$ is invariant under the first return map $T_C$. Recall that we denote by $\nu_\alpha$, the normalized counting measure supported on the period $P_\alpha \subset I_{\text{irr}}$ of the orbit of $\alpha \bmod 1$ under the Gauss map (see the last paragraph of §1.2).

**Lemma 8.7.** *Let $p : D \to I$ be the projection on the first coordinate. Then*

$$\begin{aligned}(p \circ \tau)_*(\tilde{\lambda}) &= \nu, \\ (p \circ \tau)_*(\tilde{\mu}_\alpha) &= \nu_\alpha.\end{aligned} \tag{8.11}$$

*Proof.* The first equality in (8.11) follows from the fact that $\tau_*(\tilde{\lambda}) = \lambda$ (which is basically the definition of $\tilde{\lambda}$) and the observation $p_*(\lambda) = \nu$ which was pointed out in (8.9). We argue the second equality: By Lemma 8.6 the measure $\tau_*(\tilde{\mu})$ is $\bar{S}$-invariant. By the above discussion it is finitely supported. Since $x_\alpha A$ is a loop, the first return map $T_C$ acts transitively on the support of $\tilde{\mu}$ and so the support of $\tau_*(\tilde{\mu})$ consists of a single $\bar{S}$ orbit. This clearly implies that $(p \circ \tau)_*(\tilde{\mu})$ is supported on a single periodic orbit of the Gauss map $S$. Denote this period by $P'_\alpha$. We only need to argue why $P_\alpha = P'_\alpha$, which is equivalent to $P_\alpha \cap P'_\alpha \neq \emptyset$.

Consider the matrix $g_\alpha$ defined in (2.5). The tangent vector corresponding to $g_\alpha$ defines a geodesic in $T^1\mathbb{H}$ which is a semicircle with endpoint $e_+(g_\alpha) = \alpha$. At some point along this geodesic we find a point $g$ which projects to $C^+$ under $\pi$. Let $x = \pi(g) \in C^+$ and $g' \in \mathcal{C}^+$ the corresponding point in $\mathcal{C}^+$. Clearly $x$ is in the support of $\tilde{\mu}_\alpha$ and hence the endpoint $e_+(g') = p \circ \tau(x)$ is a point of $P'_\alpha$. As the semicircle corresponding to $g_\alpha$ and the one corresponding to $g'$ are related by the action of some $\gamma \in \Gamma$ as Møbius transformation, it follows that the endpoints $\alpha, e_+(g')$ are related by the action of $\gamma$ as well. This action can effect only finitely many digits of the c.f.e of $\alpha$ and we conclude that the periods of the c.f.e of $\alpha$ and of $e_+(g')$ must be the same (up to a possible cyclic rotation) which finishes the proof. □

In light of (8.11) the following theorem clearly implies Theorem 2.1. This is mearly the corresponding statements when translated to the cross-section.

**Theorem 8.8.** *Let $\alpha, S, q,$ and $\epsilon$ be as in the statement of Theorem 2.1. For any $\kappa$-Lipschitz function $f : D \to \mathbb{C}$*

$$\left| \int_C f \circ \tau d\tilde{\mu}_{q\alpha} - \int_C f \circ \tau d\tilde{\lambda} \right| \ll_{\alpha,S,\epsilon} \max\left\{\|f\|_\infty, \kappa\right\} \operatorname{ht}(q)^{-\frac{\delta_0}{6}+\epsilon}. \tag{8.12}$$

8.5. **Relation to the equidistribution of the loops.** In light of Theorem 2.8 the content of the following lemma is clearly relevant for the proofs of Theorem 8.8 above.

**Lemma 8.9.** *Let $\alpha$ be a quadratic irrational. The suspensions $\sigma_{\tilde{\lambda}}, \sigma_{\tilde{\mu}_\alpha}$ of the probability measures $\tilde{\lambda}, \tilde{\mu}_\alpha$ are proportional to $m_X, \mu_\alpha$ respectively.*

*Proof.* The fact that $\sigma_{\tilde{\lambda}}$ is proportional to the Haar measure $m_X$ is proved in [EW11, p. 325-326]. The outline of the proof is as follows: By Lemma 8.4, $\sigma_{\tilde{\lambda}}$ is $A$-invariant. One shows that it is absolutely continuous with respect to $m_X$ and deduces the result from the ergodicity of $m_X$ with respect to the $A$-action. Regarding $\sigma_{\tilde{\mu}_\alpha}$, note that it is clearly a measure that is supported on the orbit $x_\alpha A$ and it is $A$-invariant by Lemma 8.4.

The assertion now follows from the uniqueness (up to proportionality) of an $A$-invariant measure on the periodic orbit $x_\alpha A$. □

8.6. **Concluding the proof of Theorem 8.8.** The argument yielding Theorem 8.8 is slightly technical because of the following issue: We start with a $\kappa$-Lipschitz function $f : D \to \mathbb{C}$ and construct from it the function $\widehat{f \circ \tau} : X \to \mathbb{C}$ as in Definition 8.5. As we wish to appeal to Theorem 2.8 we need to remedy $\widehat{f \circ \tau}$ to be Lipschitz in a way that will allow us to control its Lipschitz constant. In order to achieve this we shall need the following technical lemma which is proved in §9.

**Lemma 8.10.** *For any $M > 1$ and $0 < \rho < 1$ there exist a function $\varphi = \varphi_{\rho,M} : X \to [0, 1]$ with the following properties*

1. *The function $\varphi$ is $\rho^{-1}$-Lipschitz.*
2. *We have $\int_X 1 - \varphi dm_X \ll M^{-1} + \rho \log M$.*
3. *Given $f : D \to \mathbb{C}$ a $\kappa$-Lipschitz function, the product $\widehat{f \circ \tau} \cdot \varphi : X \to \mathbb{C}$ is Lipschitz with Lipschitz constant $\ll \max\{\|f\|_\infty, \kappa\} \rho^{-1} M$.*

*Proof of Theorem 8.8.* Let $\alpha, q, S, \epsilon$ be as in the statement of the theorem and $f : D \to \mathbb{C}$ a $\kappa$-Lipschitz function. Let $c_0, c_q$ be the proportionality constants satisfying $c_0 m_X = \sigma_{\tilde{\lambda}}, c_q \mu_{q\alpha} = \sigma_{\tilde{\mu}_{q\alpha}}$ whose existence is given by Lemma 8.9. Using (8.7) we have the following estimate:

$$\left|\int_C f \circ \tau d\tilde{\mu}_{q\alpha} - \int_C f \circ \tau d\tilde{\lambda}\right| = \left|\frac{c_q}{\epsilon_0}\int_X \widehat{f \circ \tau} d\mu_{q\alpha} - \frac{c_0}{\epsilon_0}\int_X \widehat{f \circ \tau} dm_X\right| \tag{8.13}$$
$$\leq \underbrace{|c_q - c_0|}_{(*)} \|f\|_\infty \epsilon_0^{-1} + c_0\epsilon_0^{-1} \underbrace{\left|\int_X \widehat{f \circ \tau} d\mu_{q\alpha} - \int_X \widehat{f \circ \tau} dm_X\right|}_{(**)}.$$

We first estimate the expression $(**)$ in (8.13). Given $M > 1, 0 < \rho < 1$ we let $\varphi = \varphi_{\rho,M}$ be as in Lemma 8.10 and denote $\psi = 1 - \varphi$.

$$(**) = \left|\int_X \widehat{f \circ \tau} \cdot (\varphi + \psi) d\mu_{q\alpha} - \int_X \widehat{f \circ \tau} \cdot (\varphi + \psi) dm_X\right| \tag{8.14}$$
$$\leq \left|\int_X \widehat{f \circ \tau} \cdot \varphi d\mu_{q\alpha} - \int_X \widehat{f \circ \tau} \cdot \varphi dm_X\right| + \left|\int_X \widehat{f \circ \tau} \cdot \psi d\mu_{q\alpha}\right| + \left|\int_X \widehat{f \circ \tau} \cdot \psi dm_X\right|.$$

We will estimate each of the three summands in the right hand side of the inequality (8.14). By Lemma 8.10(2) we have

$$\left|\int_X \widehat{f \circ \tau} \cdot \psi dm_X\right| \leq \|f\|_\infty \int_X \psi dm_X \ll \|f\|_\infty (M^{-1} + \rho \log M). \tag{8.15}$$

Next, note that by Lemma 8.10(1) $\psi$ is $\rho^{-1}$-Lipschitz and so we may apply Theorem 2.8 which together with the estimate (8.15) yields

$$\begin{aligned}\left|\int_X \widehat{f\circ\tau}\cdot\psi d\mu_{q\alpha}\right| &\leq \|f\|_\infty \int_X \psi d\mu_{q\alpha} \\ &\ll_{\alpha,S,\epsilon} \|f\|_\infty \left(\int_X \psi dm_X + \max\left\{1,\rho^{-1}\right\}\operatorname{ht}(q)^{-\frac{\delta_0}{2}+\epsilon}\right) \\ &\ll \|f\|_\infty \left(M^{-1} + \rho\log M + \rho^{-1}\operatorname{ht}(q)^{-\frac{\delta_0}{2}+\epsilon}\right). \end{aligned} \tag{8.16}$$

Finally, by Lemma 8.10(3) we can apply Theorem 2.8 to the function $\widehat{f\circ\tau}\cdot\varphi$ and conclude the following

$$\left|\int_X \widehat{f\circ\tau}\cdot\varphi d\mu_{q\alpha} - \int_X \widehat{f\circ\tau}\cdot\varphi dm_X\right| \ll_{\alpha,S,\epsilon} \max\left\{\|f\|_\infty,\kappa\right\}\rho^{-1}M\operatorname{ht}(q)^{-\frac{\delta_0}{2}+\epsilon}. \tag{8.17}$$

We now make the choice $M=\rho^{-1}=\operatorname{ht}(q)^{\frac{1}{3}(\frac{\delta_0}{2}-\epsilon)}$ and combine estimates (8.15)(8.16)(8.17) into (8.14) to obtain

$$(**) \ll_{\alpha,S,\epsilon} \max\left\{\|f\|_\infty,\kappa\right\}\operatorname{ht}(q)^{-\frac{\delta_0}{6}+\epsilon}, \tag{8.18}$$

where in the above inequality $\frac{\epsilon}{3}$ was replaced by $\epsilon$ and the term $\log(\operatorname{ht}(q))$ was absorbed in the term $\operatorname{ht}(q)^\epsilon$.

In order to finish we need to further estimate $(*)$ in (8.13). To obtain this estimation from the above we take $f: D\to\mathbb{C}$ to be identically 1 and note that in this case $\widehat{f\circ\tau}=\chi_B$ and so using (8.7) we have

$$\left|\int_X \widehat{f\circ\tau}d\mu_{q\alpha} - \int_X \widehat{f\circ\tau}dm_X\right| = \left|\mu_{q\alpha}(B)-m_X(B)\right| = \left|\frac{\epsilon_0}{c_q}-\frac{\epsilon_0}{c_0}\right|. \tag{8.19}$$

The left hand side of (8.19) is $(**)$ for this choice of $f$ and so by (8.18) we obtain

$$\left|c_q^{-1}-c_0^{-1}\right| \ll_{\alpha,S,\epsilon} \operatorname{ht}(q)^{-\frac{\delta_0}{6}+\epsilon}. \tag{8.20}$$

This establishes that $c_q\to c_0$ as $\operatorname{ht}(q)\to\infty$ and therefore in particular, for all but finitely many $q$'s $c_q>\frac{c_0}{2}$. For such $q$'s we conclude from (8.20) that

$$(*) = |c_q-c_0| \ll_{\alpha,S,\epsilon} \operatorname{ht}(q)^{-\frac{\delta_0}{6}+\epsilon}. \tag{8.21}$$

As there are only finitely many problematic $q$'s we may simply choose the implicit constant to be large enough so that (8.21) will hold for any rational $q$ supported on $S$. Note that this change in the implicit constant depends on $S$ and $\alpha$.

Plugging this estimation of $(*)$ together with (8.18) to (8.13) we obtain the desired inequality (8.12) appearing the statement of the theorem. □

*Proof of Corollary 2.5.* We use the notation introduced in the proof of Theorem 2.1 presented above. For a quadratic irrational $\alpha$ let $\tilde{P}_\alpha$ denote the support of $\tilde{\mu}_\alpha$. It follows from (8.11) that $p\circ\tau(\tilde{P}_\alpha)=P_\alpha$. It is straightforward to argue that the map $p\circ\tau:\tilde{P}_\alpha\to P_\alpha$ is always 2 to 1. The proportionality constant $c_q$ defined by $c_q\mu_{q\alpha}=\sigma_{\tilde{\mu}_{q\alpha}}$ can be rewritten as follows: The left hand side of the equation $\mu_{q\alpha}(B) = \frac{\epsilon_0}{c_q}$ (see (8.19)) can be expressed in a different way; the geodesic $x_{q\alpha}A$ which is of length $t_{q\alpha}$ penetrates $B$ exactly $\left|\tilde{P}_{q\alpha}\right|$ times and stays in $B$ along a time interval of length $\epsilon_0$ each time and so $\mu_{q\alpha}(B) = \frac{|\tilde{P}_{q\alpha}|\cdot\epsilon_0}{t_{q\alpha}}$. It follows that $c_q^{-1} = \frac{|\tilde{P}_{q\alpha}|}{t_{q\alpha}}$. By Lemma 5.1 we see that $t_{q\alpha} = c_{\mathcal{L}}(\mathrm{ht}(q))$, where the generalized branch $\mathcal{L}$ is one of the finitely many rational generalized branches $\mathcal{L}^{\tau_q}_{g_\alpha,e_f}$ (see the notation introduced in Remark 4.3(5)). We conclude from Lemma 5.1 that the ratio $c_\alpha(q) = \frac{\mathrm{ht}(q)}{t_{q\alpha}}$ attains only finitely many positive values and that if $q_n$ is a sequence such that $\mathrm{ht}(q_n|\,\mathrm{ht}(q_{n+1})$, then $c_\alpha(q_n)$ stabilizes. It now follows from (8.20) that

$$\left|c_\alpha(q)\frac{|P_{q\alpha}|}{\mathrm{ht}(q)} - c_0^{-1}\right| \ll_{\alpha,S,\epsilon} \mathrm{ht}(q)^{-\frac{\delta_0}{6}+\epsilon},$$

as desired. □

*Sketch of proof of Theorem 2.7.* In the proofs of Theorem 2.1 and Corollary 2.5 we obtained results about the c.f.e of numbers of the form $q\alpha$. The information was extracted from an understanding of the measure $\mu_{q\alpha}$ which is supported on the periodic $A$-orbit through the point on the rational generalized branch $\mathcal{L}^{\tau_q}_{g_\alpha,e_f}$ which lies on the sphere $S_{\mathrm{ht}(q)}(x_\alpha)$ in the $S$-Hecke graph $\mathcal{G}_S(x_\alpha)$. The information about the c.f.e of $q(\gamma\cdot\alpha)$ (here $\gamma\in\Gamma$ and $\gamma\cdot\alpha$ denotes the action of $\gamma$ on $\alpha$ as a Möbius transformation), is obtained by studying the periodic orbits through points on the rational generalized branches $\mathcal{L}^{\tau_q}_{g_\alpha,\gamma^{-1}}$.

The instances in the proofs of Theorem 2.1 and Corollary 2.5 in which implicit constants depending on $\alpha$ appeared, were when we appealed to Theorem 2.8 and Lemma 5.1. If instead of appealing to Theorem 2.8 we appeal to Theorem 4.9, we see that in the case $x_\alpha$ is non-split (that is, no prime in $S$ splits over $\mathbb{Q}(\alpha)$), the implicit constants may be taken independent of the generalized branch. This implies the validity of the theorem. □

## 9. Construction of $\varphi$ - Proof of Lemma 8.10

9.1. **Motivation.** We start with a function $f : D \to \mathbb{C}$ which is $\kappa$-Lipschitz and we consider the function $\tilde{f} : X \to \mathbb{C}$ given by $\tilde{f} = \widehat{f\circ\tau}$. The points of discontinuity of $\tilde{f}$ are contained in $\partial B$. We wish to find an approximation of $\tilde{f}$ which is not only continuous but for which we will have clear control on its Lipschitz constant. To achieve this, we construct an auxiliary function $\varphi$ which vanishes in an $\epsilon$-thickening of $\partial B$ and is equal to 1 outside a $2\epsilon$-thickening of $\partial B$. This will clearly make $\tilde{f}\cdot\varphi$ continuous, but in order to control its Lipschitz constant we will have to make $\varphi$ vanish 'high in the cusp' where the differential of $\tau$ explodes (see Lemma 9.6 below). Along the construction we need to pay attention to two more quantities which we should control: The Lipschitz constant of $\varphi$ and $\int\psi$, where $\psi = 1-\varphi$. These clearly fight one against the other; in order to make $\int\psi$ small we wish to take $\epsilon$ (which control the above thickening) to be small which makes the Lipschitz constant of $\varphi$ large.

Below, in §9.2-9.5, we discuss a somewhat eclectic collection of observations that we will use in order to carry out the arguments in §9.6 with little interruption.

9.2. **General metric observations.** Let $(Y, \mathrm{d})$ be a metric space. For a subset $F \subset Y$ we denote

$$(F)_\epsilon = \{y \in Y : \mathrm{d}(y, F) \le \epsilon\}\,;$$

that is, the set of all points of distance $\le \epsilon$ from $F$. The following general construction allows us to build Lipschitz functions in abundance. The proof is left to the reader.

**Lemma 9.1** (Fundamental construction)**.** *Let $(Y, \mathrm{d})$ be a metric space and $F \subset Y$ a subset. For $\epsilon > 0$ define $\varphi_{\epsilon,F} : Y \to [0, 1]$ by $\varphi_{\epsilon,F}(y) = \min\{1, \epsilon^{-1}\,\mathrm{d}(y, F)\}$. Then $\varphi_{\epsilon,F}$ attains the constant values $0$ on $F$ and $1$ on $Y \setminus (F)_\epsilon$. Furthermore, $\varphi_{\epsilon,F}$ is $\epsilon^{-1}$-Lipschitz.*

We now make two remarks regarding Lipschitz constants:

**Remark 9.2.** Consider two functions, $f : Y \to \mathbb{C}$ and $\varphi : Y \to [0, 1]$, on a metric space $(Y, \mathrm{d})$ and assume that they are $\kappa_f, \kappa_\varphi$-Lipschitz respectively with $\kappa_\varphi \ge 1$. Then, for any $x, y \in Y$ we have

$$\begin{aligned}|f \cdot \varphi(x) - f \cdot \varphi(y)| &\le |f(x) - f(y)|\,\varphi(x) + |f(y)|\,|\varphi(x) - \varphi(y)|\\ &\le 2\max\{\kappa_f, \|f\|_\infty\}\,\kappa_\varphi\,\mathrm{d}(x, y),\end{aligned}$$

that is $f \cdot \varphi$ has Lipschitz constant $\ll \max\{\kappa_f, \|f\|_\infty\}\,\kappa_\varphi$.

**Remark 9.3.** Let $f : Y \to \mathbb{C}$ be a continuous function on a metric space $(Y, \mathrm{d})$ in which between any two points $x, y$ there exists a path whose length equals $\mathrm{d}(x, y)$. Suppose there is an open cover $\{U_i\}$ of $\mathrm{supp}(f)$ such that for each $i$ the restriction $f : U_i \to \mathbb{C}$ is $\kappa$-Lipschitz. Then we claim that $f$ is $\kappa$-Lipschitz as a function on $Y$. To see this, take two points $x, y \in Y$ and connect them by a path $\gamma$ whose length is $\mathrm{d}(x, y)$. As $f$ is assumed to be continuous we can turn the open cover $\{U_i\}$ of the support of $f$ to an open cover of $Y$ by joining in the open set $U_0 = Y \setminus \mathrm{supp}(f)$. Clearly $f$ is $\kappa$-Lipschitz on $U_0$ as well. Now let $\epsilon > 0$ be a Lebesgue number for the induced open cover of the path $\gamma$. Choose points $x = x_0, x_1 \dots x_n = y$ on $\gamma$ in a monotone way (so that $\mathrm{d}(x, y) = \sum_1^n \mathrm{d}(x_i, x_{i-1})$) and such that the distance between $x_i$ to $x_{i-1}$ is less than $\epsilon$. It follows that for each $1 \le i \le n$ there exists an open set from the cover $U_{j_i}$ such that $x_{i-1}, x_i \in U_{j_i}$. As $f$ is assumed to be $\kappa$-Lipschitz on $U_{j_i}$, we conclude that

$$|f(x) - f(y)| \le \sum_1^n |f(x_i) - f(x_{i-1})| \le \sum_1^n \kappa\,\mathrm{d}(x_i, x_{i-1}) = \kappa\,\mathrm{d}(x, y).$$

9.3. **Coordinates.** We wish to define a convenient coordinate system which will allow us to carry out the relevant computations. Recall the open subsets $B, \mathcal{B}$ of of $X, G$ respectively that were defined in Lemma 8.3. We define similarly to (8.4)

$$\begin{aligned}\mathcal{B}^+ &= \left\{ga(t) : g \in \mathcal{C}^+, t \in (0, \epsilon_0)\right\} \qquad (9.1)\\ \mathcal{B}^- &= \left\{ga(t) : g \in \mathcal{C}^+, t \in (0, \epsilon_0)\right\}.\end{aligned}$$

A point $g \in \mathcal{B}$ can be written uniquely in the form $a(s)k_\theta a(t)$ where $s \in \mathbb{R}$, $t \in (0, \epsilon_0)$ and the angle $\theta \in [0, \pi)$ has some restrictions on it, arising from the requirements about the endpoints of the semicircle corresponding to $g$. We shall refer to $(s, \theta, t)$ as the *coordinates* of the point $g \in \mathcal{B}$ or of the corresponding point $\pi(g) \in B$.

As the action of $a(t)$ from the right does not effect the endpoints, the restrictions on the $\theta$-coordinate are a function of $s$ alone. We workout these restrictions for, say, $g \in \mathcal{B}^+$: We already observed (after (8.4)) that $\theta \in (\frac{\pi}{4}, \frac{\pi}{2})$ (in order to ensure that $e_+(g) \in (0,1)$). It is easy to see from the definition of the start and end points that for $s \in \mathbb{R}$, $a(s)k_\theta \in \mathcal{C}^+$, where $\theta \in [0, \pi)$, if and only if $e^s \cot\theta \in (0,1)$ and $-e^s \tan\theta < -1$. This is equivalent to saying $\tan\theta \in (\min\{e^s, e^{-s}\}, \infty)$. We choose an inverse $\tan^{-1} : \mathbb{R} \to (0, \frac{\pi}{2})$ and conclude that for a given $s$, the range of allowed angles for points $g \in \mathcal{B}^+$ with coordinates $(s, \theta, t)$, is an interval $I_s^+$ which is defined by

$$I_s^+ = (\theta_{\min}(s), \frac{\pi}{2}), \text{ where } \theta_{\min}(s) = \tan^{-1}(\min\{e^s, e^{-s}\}) > \frac{\pi}{4}. \tag{9.2}$$

Let us denote

$$\mathcal{E}^+ = \left\{(s, \theta, t) \in \mathbb{R}^3 : s \in \mathbb{R}, t \in (0, \epsilon_0), \theta \in I_s^+\right\}, \tag{9.3}$$

and define similarly $\mathcal{E}^-$ and $\mathcal{E} = \mathcal{E}^+ \cup \mathcal{E}^-$. Let $\xi : \mathbb{R}^3 \to G$ be the function

$$\xi(s, \theta, t) = a(s)k_\theta a(t). \tag{9.4}$$

Clearly, we have $\xi(\mathcal{E}) = \mathcal{B}$, $\xi(\mathcal{E}^+) = \mathcal{B}^+$, and $\xi(\mathcal{E}^-) = \mathcal{B}^-$.

**Lemma 9.4.** *There is an absolute constant $c$ such that for any $\epsilon > 0$, an $\epsilon$-ball in $\mathcal{E}$ is mapped by $\xi$ into a ball of radius $c\epsilon$ in $\mathcal{B}$.*

*Proof.* We link any two points $g_i = \xi(s_i, \theta_i, t_i) \in \mathcal{B}, i = 1, 2$ by the path which changes linearly the $s$-coordinate first, then the $\theta$-coordinate, and finally the $t$-coordinate. Each such change corresponds to the action from the right by a one-parameter subgroup $h(t)$ as in Lemma 3.3. The change in the $s$-coordinate corresponds to $h(s) = a(-t_1)k_{-\theta_1}a(s)k_{\theta_1}a(t_1)$, the change in the $\theta$-coordinate corresponds to $h(\theta) = a(-t_1)k_\theta a(t_1)$, and finally, the change in the $t$-coordinate corresponds to $h(t) = a(t)$. As the family of one-parameter subgroups that are involved in this process are conjugations of $a(t)$ and $k_\theta$, where the conjugating element is varying in a compact set, we conclude that the norm of the derivative at the identity $\dot{h}(0)$ is $\ll 1$ for some absolute implicit constant. Lemma 3.3 implies then that

$$\mathrm{d}_G(g_1, g_2) \ll |s_1 - s_2| + |\theta_1 - \theta_2| + |t_1 - t_2|,$$

which establishes the claim. □

9.4. **Height.** The map $\tau$ defined in (8.10) was considered so far as a map from the cross-section $C$. As we wish to use differentiation it will be more convenient to extend it to a map $\tau : B \to D$ in the following way: Given a point $x \in B$ it can be written uniquely as $x_C a(t)$ where $x_C \in C$ and $t \in (0, \epsilon_0)$. We define $\tau(x) = \tau(x_C)$; that is, we view $\tau$ as a function on $B$ which is constant along the direction of the geodesic flow.

As will be seen shortly, the norm of the differential of $\tau : B \to D$ is not bounded and so, in order to be able to control the Lipschitz constant of the function appearing in Lemma 8.10(3) we need to force its support to be contained in a domain in which we have some control on $\|\mathrm{d}\,\tau\|$.

Recall the Iwasawa decomposition (8.3). Let $\mathcal{F}$ denote the usual fundamental domain of $\Gamma$ in $G$, that is,

$$\mathcal{F} = \left\{ n(t)a(s)k_\theta \in G : |t| < \frac{1}{2}, t^2 + e^{2s} > 1 \right\}, \tag{9.5}$$
$$\overline{\mathcal{F}} = \left\{ n(t)a(s)k_\theta \in G : |t| \leq \frac{1}{2}, t^2 + e^{2s} \geq 1 \right\}$$

We define the height function $\mathrm{ht} : G \to \mathbb{R}$ to be $\mathrm{ht}(g) = e^s$ if $g = n(t)a(s)k_\theta$. This is indeed the imaginary coordinate of the base-point of the tangent vector to $\mathbb{H}$ corresponding to $g$. This function respects the identifications induced by $\Gamma$ on the boundary of $\overline{\mathcal{F}}$ and so descends to a function (which we continue to denote $\mathrm{ht}(\cdot)$) on $X$. For any $M > 1$ we let

$$\mathcal{H}_M = \left\{ g \in \overline{\mathcal{F}} : \mathrm{ht}(g) \geq M \right\}, \quad \mathcal{K}_M = \left\{ g \in \overline{\mathcal{F}} : \mathrm{ht}(g) < M \right\}; \tag{9.6}$$
$$H_M = \{ x \in X : \mathrm{ht}(x) \geq M \}, \quad K_M = \{ x \in X : \mathrm{ht}(x) < M \}.$$

**Remark 9.5.** It is well known that $m_X(H_M) = m_G(\mathcal{H}_M) = M^{-1}$, which is an identity that will be needed later (need to add reference).

### 9.5. Estimating norms of differentials.

**Lemma 9.6.** *The differentials of* $\tau : B \to D$ *and* $\mathrm{ht} : X \to \mathbb{R}$ *at a point* $y$ *satisfy* $\|\mathrm{d}_y\, \tau\| \ll \mathrm{ht}(y), \|\mathrm{d}_y(\mathrm{ht})\| \ll \mathrm{ht}(y)$.

*Proof.* We calculate for example $\|\mathrm{d}_y\, \tau\|$ for $y \in B^+$ (here $B^+ = \pi(\mathcal{B}^+)$). Let $N, H$, and $W$ denote the respective derivatives at time $t = 0$ of the one parameter subgroups $n(s), a(t)$, and $k_\theta$ which appear in (8.3);

$$N = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}, \; H = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \; W = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}.$$

Let $g \in \mathcal{B}^+$ be such that $y = \pi(g)$ and write $g$ as in (8.1) so that $\tau(y) = \left(\frac{a}{c}, cd\right)$ as given in (8.10). The tangent space $T_y(X)$ is identified (as an inner product space) with $T_g(G)$ which is in turn identified with the Lie algebra $\mathfrak{g} = \mathfrak{sl}_2(\mathbb{R})$ via the map sending a matrix $V \in \mathfrak{g}$ to $gV$; here we make a choice of an inner product on $\mathfrak{g}$ which induces the left-invariant Riemannnian metric on $G$ and hence on the quotient $X$. Thus, we will obtain an upper bound for the norm of $\mathrm{d}_y\, \tau$ if we calculate an upper bound for the norms in $\mathbb{R}^2$ of the vectors $\mathrm{d}_y\, \tau(gV)$ for $V = N, H, W$ (where here we abuse notation and think of $\mathrm{d}_y\, \tau$ as a map from $T_g(G)$ to $\mathbb{R}^2$).

We may think of the above $2 \times 2$ matrices as vectors in $\mathbb{R}^4$ (where the first row corresponds to the first two coordinates) and then we get that $\mathrm{d}_y\, \tau$ is given by the matrix

$$\mathrm{d}_y\, \tau = \begin{pmatrix} \frac{1}{c} & 0 & -\frac{a}{c^2} & 0 \\ 0 & 0 & d & c \end{pmatrix}.$$

A short calculation shows that

$$\mathrm{d}_y\, \tau(gN) = \begin{pmatrix} 0 \\ c^2 \end{pmatrix}, \; \mathrm{d}_y\, \tau(gH) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \; \mathrm{d}_y\, \tau(gW) = \begin{pmatrix} c^{-2} \\ c^2 - d^2 \end{pmatrix}.$$

We conclude that $\|\mathrm{d}_y\,\tau\| \ll \max\{c^2, c^{-2}, d^2\}$, where the implicit constant comes from the fact that we did not specify an inner product on $\mathfrak{g}$. Writing $g$ in its $(s,\theta,t)$-coordinates $g = a(s)k_\theta a(t)$ we calculate $c, d$ and conclude that as $|t| \le \epsilon_0$, $\|\mathrm{d}_y\,\tau\| \ll e^{|s|}$. Remark 9.8 now gives $\|\mathrm{d}_y\,\tau\| \ll \mathrm{ht}(y)$ as desired.

We briefly describe the estimate for $\mathrm{d}_y(\mathrm{ht})$. Let $g \in \overline{\mathcal{F}}$ be such that $y = \pi(g)$. Assume for a start that the Iwasawa decomposition of $g$ is given by $g = n(t)a(s)$. Then the derivative in the directions of $W$ and $N$ are trivial (because the actions from the right of the one parameter groups $k_\theta, u(t)$ do not change the height). The derivative in the direction of $H$ is $e^s$ which equals $\mathrm{ht}(y)$. It follows that for such points $\|\mathrm{d}_y(\mathrm{ht})\| \ll \mathrm{ht}(y)$. Now for the general case, let $g = n(t)a(s)k_\theta \in \overline{\mathcal{F}}$ be the Iwasawa decomposition and consider the composition $G \to G \to \mathbb{R}$ given by first acting on the right by $k_{-\theta}$ and then applying ht. As ht is invariant under the action from the right by $k_{-\theta}$, this composition equals ht. Its differential at $y$ equals by the chain rule to the composition of the differential of right multiplication by $k_{-\theta}$ at the point $y$ and the differential of ht at the point $y' = \pi(g')$, where $g' = n(t)a(s)$. As right multiplication by $k_{-\theta}$ is an isometry the first differential has norm 1 (here we use the fact that the left invariant Riemannian metric we chose on $G$ is also right $\{k_\theta\}$-invariant). We evaluated the norm of the second differential before and we conclude that the composition satisfies the desired estimate. □

**Remark 9.7.** As the differential of $\mathrm{ht} : X \to \mathbb{R}$ is $\ll M$ on $K_M$. It follows that it is Lipschitz there with a Lipschitz constant $\ll M$ (see Remark 9.9). We conclude that there exists some absolute constant $\ell$ (which is the implicit constant in the estimate $\|\mathrm{d}_y(\mathrm{ht})\| \ll \mathrm{ht}(y)$), such that the following two statements hold

(1) For any $0 < \epsilon < 1$, $(H_M)_\epsilon \subset H_{\frac{M}{\ell}}$.
(2) For any $0 < \epsilon < 1$, $(K_M)_\epsilon \subset K_{\ell M}$.

To see (1) for example, note that if this was false, then we could find $x \in K_{\frac{M}{\ell}}$ the distance of which from $H_M$ is $\le 1$. We conclude that there must be a point $x'$ such that $\mathrm{ht}(x') = M$ and $\mathrm{d}_X(x, x') \le 1$. This of course contradicts the fact that ht is $M$-Lipschitz on $K_M$.

**Remark 9.8.** We wish to comment on the height of a point $y = \pi(g) \in B$, where $g \in \mathcal{B}$ has coordinates $(s, \theta, t)$. By Lemma 3.3, if we let $g' \in \mathcal{C}$ be the point with coordinates $(s, \theta, 0)$, then $\mathrm{d}_G(g, g') \ll \epsilon_0$ (here we take $h(t) = a(t)$ to 'cancel' the $t$-coordinate in at most $\epsilon_0$ time). The height of $g'$ is by definition $\mathrm{ht}(g') = e^{|s|}$ (the reason for the absolute value is that $g'$ might be in the lower fundamental domain $k_{\frac{\pi}{2}}\mathcal{F}$). We conclude from parts (1),(2) of Remark 9.7 that

$$|s| - \log \ell \le \log(\mathrm{ht}(g)) \le |s| + \log \ell.$$

## 9.6. **The argument.**

*Proof of Lemma 8.10.* Fix $M > 1$ and $0 < \epsilon < 1$ (below $\epsilon$ replaces the number $\rho$ in the statement of Lemma 8.10). Let $F \subset X$ be defined by

$$F = (\partial B)_\epsilon \cup H_M. \tag{9.7}$$

Define $\varphi_{\epsilon,F} : X \to [0,1]$ as in Lemma 9.1. To ease the notation we simply denote it by $\varphi$ bearing in mind the dependencies on $\epsilon, M$. Lemma 9.1 implies the assertion in Lemma 8.10(1). Let $\psi = 1 - \varphi$. As $\varphi$ attains the value 1 on $X \setminus (F)_\epsilon$ we have that $\psi \leq \chi_{(F)_\epsilon}$. Furthermore, by Remark 9.7(1) and from the definitions we see that

$$(F)_\epsilon \subset (\partial B)_{2\epsilon} \cup (H_M)_\epsilon \subset ((\partial B)_{2\epsilon} \cap K_M) \cup H_{\frac{M}{\ell}}.$$

It follows that

$$\int_X \psi dm_X \leq m_X\left(((\partial B)_{2\epsilon} \cap K_M)\right) + m_X(H_{\frac{M}{\ell}}).$$

Hence, by Remark 9.5, Lemma 8.10(2) will follow once we show that the following estimate holds for all $M > 1$

$$m_X\left(((\partial B)_{2\epsilon} \cap K_M)\right) \ll \epsilon \log M. \tag{9.8}$$

In order to establish (9.8) we argue as follows: We first want to pull the calculation to $G$ and then to $\mathbb{R}^3$. It is clear that $\pi(\partial \mathcal{B} \cap \mathcal{K}_M) = \partial B \cap K_M$ and as $\pi$ can only decrease distances (that is $\pi$ is 1-Lipschitz), we must have $\pi((\partial \mathcal{B})_{2\epsilon} \cap \mathcal{K}_M) \supset (\partial B)_{2\epsilon} \cap K_M$. By the definition of the measure $m_X$ it follows that

$$m_X((\partial B)_{2\epsilon} \cap K_M) \leq m_G((\partial \mathcal{B})_{2\epsilon} \cap \mathcal{K}_M). \tag{9.9}$$

Hence, we are reduced to estimate $m_G((\partial \mathcal{B})_{2\epsilon} \cap \mathcal{K}_M)$. We will workout below the estimation for $m_G\left((\partial \mathcal{B}^+)_{2\epsilon} \cap \mathcal{K}_M\right)$ only. Let $N_\epsilon(L)$ denote the number of $\epsilon$-balls needed to cover a set $L$. Clearly,

$$N_{3\epsilon}((\partial \mathcal{B}^+)_{2\epsilon} \cap \mathcal{K}_M) \leq N_\epsilon(\partial \mathcal{B} \cap \mathcal{K}_M).$$

We know that a ball of radius $\epsilon$ in $G$ has volume $\ll \epsilon^3$ and so we deduce that

$$m_G((\partial \mathcal{B}^+)_{2\epsilon} \cap \mathcal{K}_M) \ll \epsilon^3 N_\epsilon(\partial \mathcal{B}^+ \cap \mathcal{K}_M). \tag{9.10}$$

Consider the following four subsets of $\overline{\mathcal{E}}^+ \subset \mathbb{R}^3$ which are mapped by $\xi$ onto the boundary $\partial \mathcal{B}$

$$\begin{aligned}
\mathcal{Q}_1 &= \{(s,\theta,t) : s \in \mathbb{R}, t \in (0,\epsilon_0), \theta = \theta_{\min}(s)\};\\
\mathcal{Q}_2 &= \left\{(s,\theta,t) : s \in \mathbb{R}, t \in (0,\epsilon_0), \theta = \frac{\pi}{2}\right\};\\
\mathcal{Q}_3 &= \left\{(s,\theta,t) : s \in \mathbb{R}, \theta \in I_s^+, t = 0\right\};\\
\mathcal{Q}_4 &= \left\{(s,\theta,t) : s \in \mathbb{R}, \theta \in I_s^+, t = \epsilon_0\right\}.
\end{aligned}$$

Let $\mathcal{Q} = \cup_{i=1}^4 \mathcal{Q}_i$. A point in $\mathcal{B} \cap \mathcal{K}_M$ with coordinates $(s,\theta,t)$ must satisfy $|s| \leq \log M + \log \ell$ as explained in Remark 9.8. Hence, we conclude by Lemma 9.4 that

$$N_\epsilon(\partial \mathcal{B}^+ \cap \mathcal{K}_M) \ll N_{c^{-1}\epsilon}(Q \cap \{(s,\theta,t) : |s| \leq \log M + \log \ell\}). \tag{9.11}$$

This reduces the problem to a Euclidean one: For each $1 \leq i \leq 4$ the surface

$$\mathcal{Q}_i \cap \{(s,\theta,t) : |s| \leq \log M + \log 2\}$$

is a graph of a function from a domain in $\mathbb{R}^2$ to $\mathbb{R}$. The variables vary in a range that is of bounded length in one direction and of length $2(\log M + \log \ell)$ in the other. As all these functions have derivatives which are uniformly bounded (in fact, all of them are constant apart from the function $(s,t) \mapsto \theta_{\min}(s)$ corresponding to $\mathcal{Q}_1$, see (9.2)), we deduce that

$$N_{c^{-1}\epsilon}(Q \cap \{(s,\theta,t) : |s| \le \log M + \log \ell\}) \ll \frac{\log M}{\epsilon^2}. \tag{9.12}$$

Combining (9.12),(9.11),(9.10), and (9.9) gives (9.8), which as explained above concludes the proof of Lemma 8.10(2). We turn now to the proof of Lemma 8.10(3).

Let $f : D \to \mathbb{C}$ be $\kappa$-Lipschitz and denote $\tilde{f} = \widehat{f \circ \tau}$. The support of the product $\tilde{f} \cdot \varphi$ is contained in the intersection of the supports of $\tilde{f}$ and $\varphi$. By definition of the $\widehat{\ }$ operator, the support of $\tilde{f}$ is contained in $B$. By definition of $\varphi$ its support is contained in the intersection $\{x \in X : \mathrm{d}_X(x, \partial B) \ge \epsilon\} \cap \overline{K}_M$. It follows that

$$\mathrm{supp}(\tilde{f} \cdot \varphi) \subset \{x \in B : \mathrm{d}_X(x, \partial B) \ge \epsilon\} \cap \overline{K}_M. \tag{9.13}$$

As the points of discontinuity of $\tilde{f}$ are contained in $\partial B$ we conclude that $\tilde{f} \cdot \varphi : X \to \mathbb{C}$ is continuous. In order to estimate its Lipschitz constant we wish to appeal to Remark 9.3. Cover the open set $\mathcal{B} \cap \mathcal{K}_{2M}$ by open balls $\mathcal{U}_i \subset \mathcal{B} \cap \mathcal{K}_{2M}$. Note that each $\mathcal{U}_i$ is contained in either $\mathcal{B}^+$ or $\mathcal{B}^-$. Consider the open cover $\{U_i\}$ of $\mathrm{supp}(\tilde{f} \cdot \varphi)$, where $U_i = \pi(\mathcal{U}_i)$. By Remark 9.3, Lemma 8.10(3) will follow once we prove that $\tilde{f} \cdot \varphi : U_i \to \mathbb{C}$ is $\max\{\kappa, \|f\|_\infty\} \epsilon^{-1} M$-Lipschitz. As $\varphi$ is $\epsilon^{-1}$-Lipschitz we see that by Remark 9.2 it is enough to argue that for each $i$, $\tilde{f} : U_i \to \mathbb{C}$ is Lipschitz with Lipschitz constant $\ll \kappa M$. As $U_i \subset K_{2M}$ we know by Lemma 9.6 that the norm of the differential of $\tau$ is $\ll M$ on $U_i$. It follows that the Lipschitz constant of the composition $\tilde{f} = f \circ \tau$ is $\ll \kappa M$ as desired. □

**Remark 9.9.** We remark here about a slight inaccuracy in the arguments presented above and how to remedy it: Let $M, N$ be two Riemannian manifolds and $f : U \to N$ a smooth map from an open set $U \subset M$. Assume the differential of $f$ has norm bounded by some constant $\kappa$ on $U$. We used above (in two places) the conclusion that $f$ must be $\kappa$-Lipschitz. Strictly speaking, this shows indeed that $f$ is $\kappa$-Lipschitz, but with respect to the metric induced from the restriction of the Riemannian metric from $M$ to $U$. This need not be the restricted metric on $U$ in which we are interested. In order to remedy this, one needs to prove that the following property holds: *There exists some absolute constant $c$ such that given any two points in $x, y \in U$ one is able to find a path connecting them inside $U$ of length $\le c\,\mathrm{d}(x,y)$ (here $\mathrm{d}$ is the metric of the ambient space containing $U$).*

Once this property is established, the conclusion is that $f$ has Lipschitz constant $\ll \kappa$. The above property clearly holds in any Euclidean ball. Using the fact that the exponential map from the Lie algebra to $G$ is bi-Lipschitz when restricted to a small enough neighborhood of zero, we see that any image of a small enough Euclidean ball around zero is an open neighborhood of the identity in $G$ which satisfies the desired property. Using left translations (which are isometries of $G$) we see that each point of $G$ has a basis of neighborhoods satisfying the above properties. Regarding the argument in the very end of the proof of Lemma 8.10, we should simply define the sets $\mathcal{U}_i$ to be such neighborhoods instead open balls. Regarding the use of this in Remark 9.7, we leave the details to the reader.

## References

[Arn07] V. I. Arnol′d. Continued fractions of square roots of rational numbers and their statistics. *Uspekhi Mat. Nauk*, 62(5(377)):3–14, 2007.

[Arn08] V. I. Arnol′d. Statistics of the periods of continued fractions for quadratic irrationals. *Izv. Ross. Akad. Nauk Ser. Mat.*, 72(1):3–38, 2008.

[Art82] Emil Artin. *Collected papers*. Springer-Verlag, New York, 1982. Edited by Serge Lang and John T. Tate, Reprint of the 1965 original.

[AS] Menny Aka and Uri Shapira. A soft approach to the evolution of continued fractions in a fixed quadratic field. Preprint.

[BL05] Yann Bugeaud and Florian Luca. On the period of the continued fraction expansion of $\sqrt{2^{2n+1}+1}$. *Indag. Math. (N.S.)*, 16(1):21–35, 2005.

[BO07] Yves Benoist and Hee Oh. Equidistribution of rational matrices in their conjugacy classes. *Geom. Funct. Anal.*, 17(1):1–32, 2007.

[Coh77] J. H. E. Cohn. The length of the period of the simple continued fraction of $d^{1/2}$. *Pacific J. Math.*, 71(1):21–32, 1977.

[CZ04] Pietro Corvaja and Umberto Zannier. On the rational approximations to the powers of an algebraic number: solution of two problems of Mahler and Mendès France. *Acta Math.*, 193(2):175–191, 2004.

[Duk88] W. Duke. Hyperbolic distribution problems and half-integral weight Maass forms. *Invent. Math.*, 92(1):73–90, 1988.

[ELMV] Manfred Einsiedler, Elon Lindenstrauss, Philippe Michel, and Akshay Venkatesh. Distribution of periodic torus orbits and duke's theorem for cubic fields. To appear in Annals of Mathematics.

[ELMV09] Manfred Einsiedler, Elon Lindenstrauss, Philippe Michel, and Akshay Venkatesh. Distribution of periodic torus orbits on homogeneous spaces. *Duke Math. J.*, 148(1):119–174, 2009.

[EW11] Manfred Einsiedler and Thomas Ward. *Ergodic theory with a view towards number theory*, volume 259 of *Graduate Texts in Mathematics*. Springer-Verlag London Ltd., London, 2011.

[FK10] Étienne Fouvry and Jürgen Klüners. On the negative Pell equation. *Ann. of Math. (2)*, 172(3):2035–2104, 2010.

[Gol02] E. P. Golubeva. On the class numbers of indefinite binary quadratic forms of discriminant $dp^2$. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, 286(Anal. Teor. Chisel i Teor. Funkts. 18):40–47, 227–228, 2002.

[Gri98] Guillaume Grisel. Length of continued fractions in principal quadratic fields. *Acta Arith.*, 85(1):35–49, 1998.

[Hic73] Dean R. Hickerson. Length of period simple continued fraction expansion of $\sqrt{d}$. *Pacific J. Math.*, 46:429–432, 1973.

[Kei] Matthews Keith. On the continued fraction expansion of $\sqrt{2^{2n+1}}$. Unpublished, available on http://www.numbertheory.org/pdfs/period.pdf.

[Kim03] Henry H. Kim. Functoriality for the exterior square of $\mathrm{GL}_4$ and the symmetric fourth of $\mathrm{GL}_2$. *J. Amer. Math. Soc.*, 16(1):139–183 (electronic), 2003. With appendix 1 by Dinakar Ramakrishnan and appendix 2 by Kim and Peter Sarnak.

[Lag80] J. C. Lagarias. On the computational complexity of determining the solvability or unsolvability of the equation $X^2 - DY^2 = -1$. *Trans. Amer. Math. Soc.*, 260(2):485–508, 1980.

[Ler10] Eduard Y. Lerner. About statistics of periods of continued fractions of quadratic irrationalities. *Funct. Anal. Other Math.*, 3(1):75–83, 2010.

[LW01] Elon Lindenstrauss and Barak Weiss. On sets invariant under the action of the diagonal group. *Ergodic Theory Dynam. Systems*, 21(5):1481–1500, 2001.

[Mar04] Grigoriy A. Margulis. *On some aspects of the theory of Anosov systems*. Springer Monographs in Mathematics. Springer-Verlag, Berlin, 2004. With a survey by Richard Sharp: Periodic orbits of hyperbolic flows, Translated from the Russian by Valentina Vladimirovna Szulikowska.

[McM05] Curtis T. McMullen. Minkowski's conjecture, well-rounded lattices and topological dimension. *J. Amer. Math. Soc.*, 18(3):711–734 (electronic), 2005.

[McM09] Curtis T. McMullen. Uniformly Diophantine numbers in a fixed real quadratic field. *Compos. Math.*, 145(4):827–844, 2009.

[MF93] Michel Mendès France. Remarks and problems on finite and periodic continued fractions. *Enseign. Math. (2)*, 39(3-4):249–257, 1993.

[Pol86] Mark Pollicott. Distribution of closed geodesics on the modular surface and quadratic irrationals. *Bull. Soc. Math. France*, 114(4):431–446, 1986.

[PR94] Vladimir Platonov and Andrei Rapinchuk. *Algebraic groups and number theory*, volume 139 of *Pure and Applied Mathematics*. Academic Press Inc., Boston, MA, 1994. Translated from the 1991 Russian original by Rachel Rowen.

[Ser85] Caroline Series. The modular surface and continued fractions. *J. London Math. Soc. (2)*, 31(1):69–80, 1985.

[Ven10] Akshay Venkatesh. Sparse equidistribution problems, period bounds and subconvexity. *Ann. of Math. (2)*, 172(2):989–1094, 2010.